UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística
e Computação Científica

MARCELA NUÑEZ LEMUS

ESTIMATION AND DIAGNOSTICS FOR PARTIALLY LINEAR CENSORED REGRESSION MODELS BASED ON HEAVY-TAILED DISTRIBUTIONS

Estimação e diagnóstico em modelos parcialmente lineares censurados sob distribuições de cauda pesada

CAMPINAS
2018

UNIVERSIDADE ESTADUAL DE CAMPINAS

MARCELA NUÑEZ LEMUS

ESTIMATION AND DIAGNOSTICS FOR PARTIALLY LINEAR CENSORED REGRESSION MODELS BASED ON HEAVY-TAILED DISTRIBUTIONS

Estimação e diagnóstico em modelos parcialmente lineares censurados sob distribuições de cauda pesada

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em estatística.

Dissertation presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in statistics.

Orientadora: Larissa Avila Matos

Coorientador: Víctor Hugo Lachos Dávila

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA MARCELA NUÑEZ LEMUS, E ORIENTADA PELA PROFA. DRA. LARISSA AVILA MATOS.

Agência(s) de fomento e nº(s) de processo(s): CAPES

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Nuñez Lemus, Marcela, 1989-
N922e Estimation and diagnostics for partially linear censored regression models based on heavy-tailed distributions / Marcela Nuñez Lemus. – Campinas, SP : [s.n.], 2018.

Orientador: Larissa Avila Matos.
Coorientador: Víctor Hugo Lachos Dávila.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Observações censuradas (Estatística). 2. Algoritmos de esperança-maximização. 3. Modelos lineares parciais. 4. Influência local (Estatística). 5. Misturas de escala (Estatística). I. Matos, Larissa Avila, 1987-. II. Lachos Dávila, Víctor Hugo, 1973-. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Estimação e diagnóstico em modelos parcialmente lineares censurados sob distribuições de cauda pesada
Palavras-chave em inglês:
Censored observations (Statistics)
Expectation-maximization algorithms
Partially linear models
Local influence (Statistics)
Scale mixtures
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Larissa Avila Matos [Orientador]
Guilherme Vieira Nunes Ludwing
Clécio da Silva Ferreira
Data de defesa: 23-02-2018
Programa de Pós-Graduação: Estatística

Dissertação de Mestrado defendida em 23 de fevereiro de 2018 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). LARISSA AVILA MATOS

Prof(a). Dr(a). GUILHERME VIEIRA NUNES LUDWIG

Prof(a). Dr(a). CLÉCIO DA SILVA FERREIRA

As respectivas assinaturas dos membros encontram-se na Ata de defesa

I dedicate this dissertation to God and my family.
Especially to my parents, Eduardo and Gladys,
with all my love and admiration.

Acknowledgement

* First, I am grateful to God for all the blessings received throughout my life.
* I must express my profound gratitude to my parents Eduardo and Gladys for their valuable lessons about the true meaning of life, for their sacrifice, dedication and providing me with unfailing support and continuous encouragement throughout my years of study. To my sisters Diana, Sandra and Monica for your unconditional support. Thank you!
* I want express my deepest gratitude to my advisers, professors Dra. Larissa Avila Matos and Dr. Victor Hugo Lachos for their continuous support, encouragement and patience throughout my dissertation. Their integrity and personality were a inspiring guidance for the dissertation completion. Endless thanks!
* The dissertation committee members, prof. Dr. Guilherme Vieira Nunes Ludwig and prof. Dr. Clécio Da Silva Ferreira for the contribution of many constructive suggestions toward the completion of the dissertation, which greatly improve its quality.
* I would also like to thank Christian E. Galarza for valuable helping me in my research.
* Finally to CAPES, by the financial support this entire period.

Resumo

Em muitos estudos, dados limitados ou censurados são coletados. Isso ocorre em várias situações práticas, devido as limitações dos equipamentos de medição ou pelo desenho experimental. Dessa forma, as respostas podem ser censuradas à esquerda, à direita ou em um intervalo. Por outro lado, os modelos parcialmente lineares são considerados como uma extensão flexível dos modelos de regressão lineares incluindo uma componente não paramétrica em alguma covariável. Neste trabalho, estudamos procedimentos de estimação e diagnóstico em modelos de regressão parcialmente lineares com respostas censuradas sob a classe de distribuições de mistura de escala normal (SMN). Esta família de distribuições contém um grupo de distribuições com caudas mais pesadas do que a normal que costumam ser usadas para inferências robustas de dados simétricos, como a t de Student, a slash, a normal contaminada, entre outras. Um algoritmo do tipo EM é apresentado para obter iterativamente as estimativas de máxima verossimilhança penalizada dos parâmetros dos modelos. Para examinar o desempenho dos modelos propostos, técnicas de deleção de casos e de influência local são desenvolvidas para mostrar a robustez contra observações potencialmente influentes e outliers. Isto é feito através da análise de sensibilidade das estimativas de máxima verossimilhança penalizada com alguns esquemas de perturbação no modelo ou nos dados e analisando alguns gráficos de diagnóstico. A eficácia do método proposto é avaliada através da análise de conjuntos de dados simulados e reais. O pacote `PartCensReg` implementado no R dá suporte computacional para este trabalho.

Palavras-chave: Modelo de regressão censurado, Algoritmo do tipo EM, Modelos lineares parciais, Influência local, Misturas de escala da distribuição normal.

Abstract

In many studies, limited or censored data are collected. This occurs, in many situations in practice, for reasons such as limitations of measuring instruments or due to experimental design. So, the responses can be either left, interval or right censored. On the other hand, partially linear models are considered as a flexible generalizations of linear regression models by including a nonparametric component of some covariate in the linear predictor. In this work, we discuss estimation and diagnostic procedures in partially linear censored regression models with errors following a scale mixture of normal (SMN) distributions. This family of distributions contains a group of well-known heavy-tailed distributions that are often used for robust inference of symmetrical data, such as Student-t, slash and contaminated normal, among others. A simple EM-type algorithm for iteratively computing maximum penalized likelihood (MPL) estimates of the parameters is presented. To examine the performance of the proposed model, case-deletion and local influence techniques are developed to show its robustness against outlying and influential observations. This is performed by sensitivity analysis of the maximum penalized likelihood estimates under some usual perturbation schemes, either in the model or in the data, and by inspecting some proposed diagnostic graphs. We evaluate the finite sample performance of the algorithm and the asymptotic properties of the MPL estimates through empirical experiments. An application to a real dataset is presented to illustrate the effectiveness of the proposed methods. The package `PartCensReg` implemented for the software R give computational support to this work.

Keywords: Censored regression model, EM-type algorithm, Partially linear models, Local influence, Scale mixtures of normal distributions.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The problem of estimation of a regression model where the dependent variable is censored has been studied in different fields, such as econometric analysis and clinical testing, among many others. For example, in AIDS research, the viral load measures may be subject to some lower and upper detection limits, below or above which they are not quantifiable. As a result, the viral load responses are either left or right censored depending on the diagnostic assays used (see, for instance, Wu, 2010).

In the framework of censored regression (CR) models, the random errors are routinely assumed to follow a normal distribution for mathematical convenience. However, if the random error distribution is non-normal, in particular, if its tails are heavier than normal ones, then the accuracy of the ordinary least squares solutions is lost, introducing biases in the parameter estimates. For more accurate models, a large number of parametric models to extend well-known distributions and to provide flexibility in modeling data have been investigated in recent years. For instance, Arellano-Valle et al. (2012) advocated the use of the Student-t distribution in the context of CR models. More recently, Massuia et al. (2015) developed diagnostic measures for CR models using the Student-t distribution, including the implementation of an interesting (and simple) expectation-maximization (EM) algorithm for maximum likelihood (ML) estimation. Garay et al. (2015, 2017) proposed a CR model with observational errors following a SMN distribution (SMN-CR model) from Bayesian and likelihood based perspectives, respectively. They demonstrated the robustness of the SMN-CR model against outliers through extensive simulations.

Partially linear regression (PLR) models belong to the class of semiparametric regression models (see, for instance, Härdle et al., 2004). They are quite flexible since the nonparametric component can model nonlinear behavior introduced by some covariate in the model. Linear regression models can be seen as a limiting case of PLR models when the nonparametric

component is not considered. Comprehensive surveys are available in Green and Silverman (1993) and Härdle et al. (2004). In the past few years, several works on PLR under flexible error distributions have been published. For instance, Ibacache-Pulgar et al. (2013) developed diagnostic measures for PLR models using the Student-t distribution, Relvas and Paula (2016) derived an iterative estimation process and some diagnostic procedures in PLR with AR(1) symmetrical errors. Ferreira and Paula (2017) proposed a PLR model allowing the errors to follow a skew-normal (Azzalini, 1985) distribution. In the context of partial linear censored regression (PCR) models, Vanegas and Paula (2017) proposed the log-symmetric regression model, where the presence of non-informative censored observations is admitted. Castro et al. (2014) advocated the use of the SMN class of distributions in PCR (SMN-PCR) models and adopted a Bayesian framework to carry out posterior inference.

Since the classic normal model is very sensitive to outlying observations, the assessment of robustness of the parameter estimates is an important concern. The deletion method, which consists of studying the impact on the parameter estimates after dropping individual observations, is probably the most employed technique to detect influential observations (see Cook and Weisberg (1982) and the references therein). Nevertheless, research on the influence of small perturbations in the model(or data) on the parameter estimates has received increasing attention in recent years. This can be achieved by performing local influence analysis, a general statistical technique used to assess the stability of the estimation outputs with respect to the model inputs. This research area has received considerable attention in the statistical literature for linear regression models since the seminal work of Cook (1986). However, for the SMN-PCR model, the marginal log-likelihood function is too complex for many applications and a direct application of Cook's approach may be cumbersome, since first and second partial derivatives of this function are involved. Zhu and Lee (2001) presented an approach to perform local influence analysis for general statistical models with missing (or incomplete) data by working with a Q-displacement function, closely related to the conditional expectation of the complete-data log-likelihood used at the *E-step* of the ECME algorithm. This approach produces results very similar to those obtained from Cook's method. Moreover, case-deletion can be also studied by using the Q-displacement function following the approach of Zhu et al. (2001). These methods, and their variants, have been applied successfully to perform influence analysis in several CR models, as seen in Matos et al. (2013) and Massuia et al. (2015), among others. In this work, we develop a local influence approach using this method for the SMN-PCR model, showing that it leads to simple influence measures. The proposed estimation and diagnostic method are implemented in the R package `PartCensReg` (Lemus et al., 2018) available in the CRAN repository.

Although some works in PCR models with symmetrical distributions have been recently published, so far, to the best of our knowledge, there is no attempt on studying the SMN-PCR model from a likelihood based perspective. The goals of this dissertation is develop a fully likelihood-based approach for computing the maximum penalized likelihood (MPL) estimates of the parameters through of an efficient EM-type algorithm (the ECME algorithm) and also propose diagnostic tools under this model.

1.1 Preliminaries

We begin by defining some notation and presenting the basic concepts which are used throughout this work. A normal distribution with mean μ and variance σ^2 is denoted by $\text{N}(\mu, \sigma^2)$, where $\phi\left(\cdot|\mu, \sigma^2\right)$ denotes its probability density function (pdf). Also, $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the pdf and the cumulative distribution function (cdf) of the standard normal distribution. On the other hand, $F_{SMN}(.)$, $F_{PVII}(\cdot)$, $F_{SL}(\cdot)$ and $F_{CN}(\cdot)$ represent the cdf of the standard SMN, standard slash, standard Pearson type VII and the standard contaminated normal distributions, respectively. When a random variable X follows a Gamma(a, b) distribution, we will consider the shape-rate parameterization, i.e., with mean a/b and variance a/b^2, where $a > 0$ and $b > 0$. We use the traditional convention of denoting a random variable (or a random vector) by an upper-case letter and its realization by the corresponding lower-case letter. Random vectors and matrices are denoted by boldface letters.

Definition 1. A random variable Y is said to have a SMN distribution with location parameter $\mu \in \mathcal{R}$, scale parameter $\sigma^2 \in (0, \infty)$ and an auxiliary vector of parameters $\boldsymbol{\nu} \in \mathcal{R}^k$, denoted by $Y \sim$ SMN$(\mu, \sigma^2, \boldsymbol{\nu})$, if it has the following stochastic representation:

$$Y \stackrel{d}{=} \mu + k(U)^{1/2} Z, \tag{1.1.1}$$

where Z and U are independent random variables, $Z \sim \text{N}(0, \sigma^2)$, $k(\cdot)$ is a positive weight function, U is a mixing positive random variable with cdf $H(\cdot|\boldsymbol{\nu})$, with $\boldsymbol{\nu}$ being a scalar or vector parameter indexing the distribution of U and $\stackrel{d}{=}$ means "has the same distribution as". It is easy to see from (1.1.1) that $Y|k(U) = k(u) \sim \text{N}(\mu, k(u)\sigma^2)$. Using conditional distribution, the pdf of (1.1.1) is:

$$f_{SMN}(y|\mu, \sigma^2, \boldsymbol{\nu}) = (2\pi\sigma^2)^{-1/2} \int_0^\infty k(u)^{-1/2} \text{exp}\left[-k^{-1}(u)\frac{(\text{y} - \mu)^2}{2\sigma^2}\right] dH(u|\boldsymbol{\nu}). \tag{1.1.2}$$

When it is considered $k(U) = U^{-1}$, the SMN class of distributions include as particular cases the Student-t, slash, normal contaminated distributions, among others. Also include the normal distribution as a special case (Andrews and Mallows, 1974).

- If U is degenerated in 1, i.e. $\mathrm{P}(U=1)=1$, then $Y \sim \mathrm{N}(\mu, \sigma^2)$.

- If $U \sim \mathrm{Gamma}(\nu/2, \nu/2)$, Y follows a Student-t distribution with $\nu > 0$, then we have that $Y \sim \mathrm{T}(\mu, \sigma^2, \nu)$ and its pdf is

$$f_T(y|\mu,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu\sigma^2}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{d(y)^2}{\nu}\right)^{\frac{-\nu+1}{2}}, \qquad -\infty < y < \infty,$$

  with $d(y) = \frac{y-\mu}{\sigma}$. The Student-t distribution reduces to the normal distribution when $\nu \uparrow \infty$. If $\nu = 1$, we have as a particular case the Cauchy distribution.

- If $U \sim \mathrm{Beta}(\nu, 1)$, Y follows a slash distribution with location parameter $\mu \in \mathcal{R}$, scale $\sigma^2 \in (0, \infty)$ and shape $\nu > 0$, denoted by $Y \sim \mathrm{SL}(\mu, \sigma^2, \nu)$. The pdf is given by

$$f_{SL}(y|\mu,\sigma^2,\nu) = \nu\int_0^1 u^{\nu-1}\phi(y;\mu,u^{-1}\sigma^2)d(u), \qquad -\infty < y < \infty.$$

  When $\nu \uparrow \infty$, the slash distribution reduces to the normal distribution.

- If U is a discrete random variable that assumes the following values

$$U = \begin{cases} \gamma & \text{with probability} \quad \varphi; \\ 1 & \text{with probability} \quad (1-\varphi), \end{cases}$$

  the associated density of U is

$$h(u, \boldsymbol{\nu}) = \varphi I_{(u=\gamma)} + (1-\varphi) I_{(u=1)}.$$

  And, the pdf of Y takes the form of

$$f_{CN}(y|\mu,\sigma^2,\boldsymbol{\nu}) = \varphi\phi(y;\mu,\gamma^{-1}\sigma^2) + (1-\varphi)\phi(y;\mu,\sigma^2), \qquad -\infty < y < \infty.$$

  So, we will denote by $Y \sim \mathrm{CN}(\mu, \sigma^2, \boldsymbol{\nu})$, if Y follows a contaminated normal distribution, with $\boldsymbol{\nu}^\top = (\varphi, \gamma)^\top$, $0 \leq \varphi \leq 1$ and $0 < \gamma \leq 1$. The parameter φ can be interpreted as the proportion of outliers and γ as a scale factor (Lachos et al., 2011). In this case, if $\gamma = 1$, we have the normal distribution.

The EM algorithm and some of its extensions

Introduced by Dempster et al. (1977), the EM algorithm (Expectation–Maximization) it is an optimization method widely used to obtain maximum likelihood estimates when there is presence of missing data, censoring, and/or latent variables. The fundamental idea is consider

the representation of a model in which the observations are augmented in a structure of latent observations based on a stochastic representation, this is, in terms of simpler distributions that depend on unobservable quantities. The main characteristics of the EM algorithm is the ease implementation in computational terms and the monotone convergence (McLachlan and Krishnan, 2008). The algorithm maximizes the complete log-likelihood function $\ell_c(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}}, \boldsymbol{z})$ at each step, where $\boldsymbol{z}$ is the vector that contains all the information including the latent observations and under mild regularity conditions converging quickly to a stationary point of the observed log-likelihood denoted by $\ell(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})$ (Wu, 1983). The algorithm consists of two steps

- ***E-step***: Replace the observed log-likelihood by the complete log-likelihood and compute the conditional expectation $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \text{E}_{\boldsymbol{\theta}^{(k)}}[\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z})|\mathbf{y}_{\text{obs}}, \widehat{\boldsymbol{\theta}}^{(k)}]$, where $\widehat{\boldsymbol{\theta}}^{(k)}$ is the estimate obtained in the k-th iteration.

- ***M-step***: Maximize the function $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ to obtain $\widehat{\boldsymbol{\theta}}^{(k+1)}$ such that $\widehat{\boldsymbol{\theta}}^{(k+1)} = \underset{\boldsymbol{\theta}\in\Theta}{arg\,max}\{Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})\}$.

In some situations, from the numerical point of view, maximize simultaneously all the components of the vector $\boldsymbol{\theta}$ is difficult, the complete-log likelihood itself may be complicated. Meng and Rubin (1993) generalized the EM algorithm to the ECM algorithm. Here the *M-step* is replaced with a set of conditional maximization steps, where the parameter vector is partitioned into several subsets and the estimate is based on a conditional set of all the others. However in some models, such as the one developed in this dissertation, it is convenient to use an extension of the EM and ECM algorithms, the expectation-conditional maximize either (ECME) algorithm (Liu and Rubin, 1994), where some or all *CM-step* of the ECM algorithm are replaced by some steps that conditionally maximize the incomplete-data log likelihood function and not the Q-function, i.e. each *CM-step* maximizes the conditional expectation of the complete-data log likelihood and in others maximize the constrained actual marginal likelihood function (*CML-step*). The convergence results hold for the ECM and ECME algorithms. Finally, the algorithm is iterated until a certain convergence criterion is small enough, like successive evaluations of the actual log-likelihood: $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})||$ or $||\ell(\boldsymbol{\theta}^{(k+1)})/\ell(\boldsymbol{\theta}^{(k)}) - 1||$.

1.2 Organization of the Dissertation

The dissertation is divided into four chapters and an appendix. In the present chapter we briey discussed some preliminary results related to SMN class of distributions as well as important results for the development of our proposed EM-type algorithm. In Chapter

2 the SMN-PCR model is defined and the maximum penalized likelihood (MPL) estimation procedure based in the EM-type algorithm is presented. We also introduce the procedure to obtain the approximated standard errors via the observed information matrix and the influence diagnostic techniques, considering case-deletion and local influence approaches. In Chapter 3 numerical examples using both simulated and real datasets are presented to evaluate and illustrate the performance of the proposed methodology. Finally, in Chapter 4 we will provide some conclusions remarks, with some recommendations for future research.

Chapter 2

The SMN-PCR model and diagnostic analysis

Many data problems requires techniques goes beyond simple linear regression. Semiparametric regression models are statistical models that allow the mean response of interest to be linearly dependent on some explanatory variables and in other variable it not. In general, we assume a decomposition of the explanatory variables in two components, the first one consisting of continuous and categorical explanatory variables that influence the response variable linearly, while the other component is characterized by a nonparametric function (continuous variables). Therefore, these models allow the incorporation of important characteristics of the data, such as the nonlinearity of some variables. In particular, we consider the partially linear regression (PR) model $Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(t_i) + \varepsilon_i$, where $\boldsymbol{\beta}$ and $\mathbf{f}(\cdot)$ are the regression parameters and the smooth function of the auxiliary variable t to be estimated respectively and ε_i denote an error term with mean zero and constant variance. The estimation of $\boldsymbol{\beta}$ and $\mathbf{f}(\cdot)$ has been studied in different contexts as smoothing splines (Heckman, 1986), kernel smoothing (Speckman, 1988) and penalized splines (Green and Silverman, 1993; Ruppert et al., 2003; Liang, 2006; Holland, 2017).

The data, in some situations, exhibit some important features should be considered. The problem information loss often occurs in many scientific fields, such as environmental sciences, biomedical and engineering, among others, where only partial information of some observations is reported (Wu, 2010). These partial information is due to limits of quantification of the assay used, that is, the real values of the observations may be bellow or above these limits. These types of observations are known as censored data. The solution used by some practitioners for this issue is disregard censored cases or replace these observations for the values associated with the detection limits and consequently the results obtained will be biased. Besides, the presence of atypical and influential observations is common and it is suggested in the literature

use a more flexible class of distributions that allow to accommodate these kind of observations, such as the scale mixtures of normal (SMN) distributions. In this Chapter, we have developed a complete methodology for partially linear regression models with censored data under the SMN distributions through of an efficient ECME algorithm for the maximum penalized likelihood (MPL) estimates and diagnostics measures.

2.1 The model

Let us consider a partially linear model, where the responses $Y_1, \ldots, Y_n$ are random variables with independent and identically distributed errors according to a SMN distribution. To be more precise, let us write:

$$\begin{aligned} Y_i &= \mathbf{x}_i^\top \boldsymbol{\beta} + f(t_i) + \varepsilon_i, \\ \varepsilon_i &\overset{\text{iid.}}{\sim} \text{SMN}(0, \sigma^2, \boldsymbol{\nu}), \end{aligned} \tag{2.1.1}$$

where $i = 1, \ldots, n$, Y_i is the response for subject i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of regression parameters of dimension $p \times 1$; $\mathbf{x}_i = (x_{1_i}, \ldots, x_{p_i})^\top$ is a $p \times 1$ vector of explanatory variable values, t_i is a scalar that may represent a value of a continuous variable, for example time and $f(\cdot)$ is a smooth function. We have that $\mathbf{x}_i$ it's not colinear with $f(\cdot)$.

In this work, we are interested in censored observations in the response variable, i.e observations for which the value assumed for Y_i is not known, however it is deduced that its value belongs to the interval $\mathscr{T}$. So, if the response variable is right-censored we have $\mathscr{T} = [\tau, \infty)$ and it is left-censored, $\mathscr{T} = [-\infty, \tau)$. Therefore, an indicator variable V_i is also observed, which assumes the value of 1 when Y_i it is censored and 0 when it is not censored. In the case we consider left-censored observations

$$Y_{\text{obs}_i} = \begin{cases} \tau_i & \text{if} \quad Y_i \leq \tau_i; \\ Y_i & \text{if} \quad Y_i > \tau_i, \end{cases} \tag{2.1.2}$$

$i = 1, \ldots, n$. We have chosen to work with the left censored case, but the results are easily extendable to other censoring types. We call the model defined in Equations (2.1.1)-(2.1.2) the SMN-PCR model. Alternatively, the model (2.1.1) can be written as:

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{n}_i^\top \mathbf{f} + \varepsilon_i, \tag{2.1.3}$$

where $\mathbf{f} = (f(t_1^0), \ldots, f(t_r^0))^\top$ is an $r \times 1$ vector with $t_1^0, \ldots, t_r^0$ being the distinct and ordered values of t_i; $\mathbf{n}_i$ is an $r \times 1$ incidence vector with the s-th element equal to the indicator function

$I(t_i = t_s^0)$ for $s = 1, \ldots, r$. In matrix form, model (2.1.3), can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{2.1.4}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is the response vector of dimension $n \times 1$, $\mathbf{X}$ is an $n \times p$ design matrix, $\mathbf{N}$ is an $n \times r$ incidence matrix with the (i, s)-th element equal to the indicator function $I(t_i = t_s^0)$, for $s = 1, \ldots, r$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ is an $n \times 1$ vector of random errors with elements belonging to the SMN class of distributions.

2.1.1 The log-likelihood function

Given an observed sample $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_n)^\top$ of $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, where there are m censored values of the characteristic of interest, we can partition the observed sample $\mathbf{y}_{\text{obs}}$ into two subsamples of m censored and $n - m$ uncensored values, such that $\mathbf{y}_{\text{obs}} = \{\tau_1, \ldots, \tau_m, y_{m+1}, \ldots, y_n\}$. Then, the log-likelihood function of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{f}^\top, \sigma^2, \boldsymbol{\nu}^\top)^\top$ considering left-censored observation is given by:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log\left[\prod_{i=1}^{n}\left[F_{SMN}\left(\frac{\tau_i - \mu_i}{\sigma}\right)\right]^{I_i}\left[f_{SMN}(y_i|\mu_i, \sigma^2, \boldsymbol{\nu})\right]^{1-I_i}\right], \\ &= \sum_{i=1}^{m}\log\left[F_{SMN}\left(\frac{\tau_i - \mu_i}{\sigma}\right)\right] + \sum_{i=m+1}^{n}\log\left[f_{SMN}(y_i|\mu_i, \sigma^2, \boldsymbol{\nu})\right], \end{aligned} \tag{2.1.5}$$

where $\mu_i = \mathbf{x}_i^\top\boldsymbol{\beta} + \mathbf{n}_i^\top\mathbf{f}$; $I_i = 1$ if $y_i \leq \tau_i$ and $I_i = 0$ otherwise. Maximization of Equation (2.1.5) without imposing restrictions over the function $\mathbf{f}(\cdot)$ may cause over-fitting and non-identification of $\boldsymbol{\beta}$ (see, for instance, Green, 1987). A well-known procedure is based on the penalized log-likelihood, which consists of incorporating a penalty function in the log-likelihood function, such that:

$$\ell_p(\boldsymbol{\theta}, \alpha) = \ell(\boldsymbol{\theta}) - \frac{\alpha}{2}J(\mathbf{f}), \tag{2.1.6}$$

where $\ell_p(\boldsymbol{\theta}, \alpha)$ denotes the penalized log-likelihood function, $J(\mathbf{f})$ is the penalty function over $\mathbf{f}(\cdot)$ and α is a smoothing parameter that controls the tradeoff between goodness-of-fit and the estimated function's smoothness. The MPL estimates are obtained by maximizing the penalized log-likelihood defined in (2.1.6). The MPL estimation problem based on an efficient ECME algorithm is considered in the next section.

2.1.2 Parameter estimation via an ECME algorithm

To implement the EM method, we require a representation of the model in terms of missing data. First, observe that by Equation (1.1.1) and given $U_i = u_i$, if $Y_i \sim \text{SMN}(\mu_i, \sigma^2, \boldsymbol{\nu})$ then:

$$\begin{aligned} Y_i|U_i &= u_i \sim \text{N}(\mu_i, u_i^{-1}\sigma^2), \\ U_i &\sim \text{H}(\cdot|\boldsymbol{\nu}). \end{aligned} \tag{2.1.7}$$

This relationship is a convenient hierarchical representation of the SMN-PCR model, and will be useful in *E-step* of the algorithm. The key to the development of our ECME algorithm is to consider the augmented dataset $\boldsymbol{z} = \{\tau_1, \ldots, \tau_m, y_{m+1}, \ldots, y_n, u_1, \ldots, u_n\}$. As a consequence, we can use the representation in (2.1.7) to obtain the complete-data penalized log-likelihood, given by:

$$\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z}) = -\frac{n}{2}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^{n}\log u_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}u_i(y_i-\mu_i)^2 + \sum_{i=1}^{n}\log h(u_i|\nu) - \frac{\alpha}{2}J(\mathbf{f}) + C,$$

where $\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z})$ is the complete penalized log-likelihood function, $h(\cdot|\nu)$ is the density of the mixing variable U and C is a constant independent of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{f}^\top, \sigma^2)^\top$. Like Ibacache-Pulgar et al. (2013) and Ferreira and Paula (2017), we consider the following penalty function:

$$J(\mathbf{f}) = \int_a^b [f''(t)]^2 dt,$$

where $[f''(t)]$ denotes the second derivative of $f(t)$ with [a, b] containing the values t_j^0, for all $j = 1, \ldots, r$. As in Green and Silverman (1993), we use the natural cubic spline as a solution for the smoothing function $f(\cdot)$, therefore $J(\mathbf{f}) = \mathbf{f}^\top\mathbf{K}\mathbf{f}$, where $\mathbf{K} \in \mathcal{R}^{r\times r}$ is a non-negative definite matrix that depends only on the knot differences (see also Annex I). A complete expression of $\mathbf{K}$ may be found, for instance, in Green and Silverman (1993).

In the *E-step* of the algorithm, we must obtain the so-called Q-function,

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \text{E}_{\boldsymbol{\theta}^{(k)}}[\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z})|\mathbf{y}_{\text{obs}}, \widehat{\boldsymbol{\theta}}^{(k)}],$$

in which the superscript (k) indicates the estimate of the related parameter at stage k of the algorithm and $\text{E}_{\boldsymbol{\theta}^{(k)}}$ is the conditional expectation of the complete penalized log-likelihood function given the current estimate $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$. Thus, dropping the constants and given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$,

the Q-function can be written as:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \propto -\frac{n}{2}\log\widehat{\sigma^2}^{(k)} - \frac{1}{2\widehat{\sigma^2}^{(k)}}\sum_{i=1}^{n}\left[\xi_{2_i}(\widehat{\boldsymbol{\theta}}^{(k)}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}}^{(k)})\widehat{\mu}_i^{(k)} + \xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})\widehat{\mu}_i^{(k)^2}\right] - \frac{\alpha}{2}\widehat{\mathbf{f}}^{(k)^\top}\mathbf{K}\widehat{\mathbf{f}}^{(k)} \tag{2.1.8}$$

$$\propto -\frac{n}{2}\log\widehat{\sigma^2}^{(k)} - \frac{1}{2\widehat{\sigma^2}^{(k)}}(\mathbf{1}_n^\top\boldsymbol{\xi}_2^{(k)} - 2\widehat{\boldsymbol{\mu}}^{(k)\top}\boldsymbol{\xi}_1^{(k)} + \widehat{\boldsymbol{\mu}}^{(k)\top}\boldsymbol{\Omega}^{(k)}\widehat{\boldsymbol{\mu}}^{(k)}) - \frac{\alpha}{2}\widehat{\mathbf{f}}^{(k)^\top}\mathbf{K}\widehat{\mathbf{f}}^{(k)}.$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with elements $\xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})$ of dimension $n \times n$, $\boldsymbol{\xi}_1^{(k)}$ $\boldsymbol{\xi}_2^{(k)}$ are vectors of dimension $n \times 1$ with elements $\xi_{1_i}(\widehat{\boldsymbol{\theta}}^{(k)})$, $\xi_{2_i}(\widehat{\boldsymbol{\theta}}^{(k)})$ respectively, $\widehat{\boldsymbol{\mu}}^{(k+1)}$ is the $n \times 1$ vector of means at the k-th iteration and $\mathbf{1}_n$ a $(n \times 1)$ vector of ones.

Therefore, it is clear that the expression of the Q-function depends completely on the knowledge of the expectations

$$\xi_{s_i}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | y_{\text{obs}_i}], \quad s = 0, 1, 2.$$

Now, observe that

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_{\text{obs}_i}] &= \mathrm{E}_{\boldsymbol{\theta}^{(k)}}\left\{\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_i] | y_{\text{obs}_i}\right\} \\ &= \mathrm{E}_{\boldsymbol{\theta}^{(k)}}\left\{Y_i^s \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i | Y_i] | y_{\text{obs}_i}\right\}.\end{aligned}$$

The following result is very important for the development of our proposed ECME algorithm. It was provided and proved by Garay et al. (2017, Proposition 1), and is an extension of Theorem 1 and Corollary 1 in Genç (2013). Let $Y \sim \text{SMN}(0, 1, \boldsymbol{\nu})$ with scale factor U and mixture distribution $H(\cdot|\boldsymbol{\nu})$. Thus for $a < b$, $E[U^r Y^s | Y \in \mathscr{A}]$ for $r \geq 1$, $\mathscr{A} = (a, b)$ and $s = 0, 1, 2$ are given by:

$$E[U^r | Y \in \mathscr{A}] = \zeta(a, b)\left[E_\Phi(r, b) - E_\Phi(r, a)\right], \tag{2.1.9}$$

$$E[U^r Y | Y \in \mathscr{A}] = \zeta(a, b)\left[E_\phi(r - 0.5, a) - E_\phi(r - 0.5, b)\right], \tag{2.1.10}$$

$$E\left[U^r Y^2 | Y \in \mathscr{A}\right] = \zeta(a, b)\left[E_\Phi(r - 1, b) - E_\Phi(r - 1, a) + aE_\phi(r - 0.5, a - bE_\phi(r - 0.5, b)\right], \tag{2.1.11}$$

with $\zeta(a, b) = (F_{SMN}(b) - F_{SMN}(a))^{-1}$ and

$$\begin{aligned}E_\phi(r, h) &= E\left[U^r \phi\left(h\, U^{0.5}\right)\right] = \int_0^\infty u^r \phi\left(h\, u^{0.5}\right) dH(u|\boldsymbol{\nu}), \\ E_\Phi(r, h) &= E\left[U^r \Phi\left(h\, U^{0.5}\right)\right] = \int_0^\infty u^r \Phi\left(h\, u^{0.5}\right) dH(u|\boldsymbol{\nu}).\end{aligned}$$

As was previously mentioned, the SMN class of distributions include as particular cases the Student-t, slash, contaminated normal and the normal distribution, among others. Therefore, the calculation of

$E_\phi(r,h)$ and $E_\Phi(r,h)$ will depend on the type of distribution (see Table 2.1). We refer to Garay et al. (2017) for proofs and additional properties. For a *censored observation* i, $Y_i \leq \tau_i$, we have

$$\xi_{s_i}(\boldsymbol{\theta}^{(k)}) = \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i Y_i^s | Y_i \leq \tau_i], \tag{2.1.12}$$

which was obtained in Garay et al. (2017, Proposition 1) with expression given by the Equations (2.1.9) to (2.1.11). Table 2.1 presents the expressions for some members of the family SMN.

Table 2.1: $E_\phi(r,h)$ and $E_\Phi(r,h)$ for some members of the SMN family of distributions (Garay et al., 2017).

Distribution	$E_\phi(r,h)$	$E_\Phi(r,h)$
Student-t	$\frac{\Gamma\left(\frac{\nu+2r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi}}\left(\frac{\nu}{2}\right)^{\nu/2}\left(\frac{h^2+\nu}{2}\right)^{-\frac{(\nu+2r)}{2}}$	$\frac{\Gamma\left(\frac{\nu+2r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu}{2}\right)^{-r} F_{PVII}(h\|\nu+2r,\nu)$
Slash	$\frac{\nu}{\sqrt{2\pi}}\left(\frac{h^2}{2}\right)^{-(\nu+1)}\Gamma\left(\nu+r, 0.5h^2\right)$	$\left(\frac{\nu}{\nu+1}\right) F_{SL}(h\|\nu+r)$
Contaminated normal	$\varphi\gamma^r\phi(h\sqrt{\gamma}) + (1-\varphi)\phi(h)$	$\gamma^r F_{CN}(h\|\varphi,\gamma) + (1-\gamma^r)\Phi(h)$

On the other hand, for *an uncensored observation* i, we have

$$\xi_{s_i}(\boldsymbol{\theta}^{(k)}) = y_i^s \mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i | Y_i]. \tag{2.1.13}$$

The values of $\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i|Y_i]$ were computed before by Osorio et al. (2007) and are presented in Table 2.2, with $d(\boldsymbol{\theta}^{(k)}, y_i) = (y_i - \mu_i^{(k)})/\sigma^{(k)}$. Here, $\boldsymbol{\xi}_s^{(k)}$ will denote the vector containing the $\xi_{s_i}(\widehat{\boldsymbol{\theta}}^{(k)})$ elements in which, if the observation i is censored, it will be computed using $\xi_{s_i}(\widehat{\boldsymbol{\theta}}^{(k)})$ given in (2.1.12), or else using $\xi_{s_i}(\widehat{\boldsymbol{\theta}}^{(k)})$ as in (2.1.13), for $s = 0, 1, 2$. Note that, $\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[\log h(U_i|\boldsymbol{\nu})|y_{\mathrm{obs}_i}]$ and $\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[\log(U_i)|y_{\mathrm{obs}_i}]$ depend only on $\boldsymbol{\nu}$, which is assumed known at this stage.

Table 2.2: $\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i|Y_i]$ for some members of the SMN family of distributions.

	Distribution		
	Student-t	Slash	Contaminated normal
$\mathrm{E}_{\boldsymbol{\theta}^{(k)}}[U_i\|y_i]$	$\frac{(\nu+1)}{\nu + d^2(\boldsymbol{\theta}^{(k)}, y_i)}$	$\frac{\Gamma\left(\nu+1.5, d^2(\boldsymbol{\theta}^{(k)}, y_i)/2\right)}{\Gamma\left(\nu+0.5, d^2(\boldsymbol{\theta}^{(k)}, y_i)/2\right)}$	$\frac{1-\varphi+\varphi\gamma^{1.5}e^{0.5(1-\gamma)d^2(\boldsymbol{\theta}^{(k)}, y_i)}}{1-\varphi+\varphi\gamma^{0.5}e^{0.5(1-\gamma)d^2(\boldsymbol{\theta}^{(k)}, y_i)}}$

Thus, the proposed ECME algorithm can be summarized in the following steps:

1. ***E-step***: Given $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, compute $\xi_{s_i}(\widehat{\boldsymbol{\theta}}^{(k)})$ or $\boldsymbol{\xi}_s$ in matrix form, for $s = 0, 1, 2$.

2. ***CM-step***: Update $\widehat{\boldsymbol{\theta}}^{(k)}$ by maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ over $\boldsymbol{\theta}$, which leads to the following expressions

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(k+1)} &= \left[\sum_{i=1}^{n} \xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})\mathbf{x}_i\mathbf{x}_i^\top\right]^{-1} \sum_{i=1}^{n} \mathbf{x}_i \left[\xi_{1_i}(\widehat{\boldsymbol{\theta}}^{(k)}) - \xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})\mathbf{n}_i^\top \widehat{\mathbf{f}}^{(k)}\right] \\
&= \left(\mathbf{X}^\top \boldsymbol{\Omega}^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \left(\boldsymbol{\xi}_1^{(k)} - \boldsymbol{\Omega}^{(k)} \mathbf{N}^\top \widehat{\mathbf{f}}^{(k)}\right), \\
\widehat{\mathbf{f}}^{(k+1)} &= \left[\sum_{i=1}^{n} \xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})\mathbf{n}_i\mathbf{n}_i^\top + \widehat{\alpha}^{(k)}\widehat{\sigma^2}^{(k)}\mathbf{K}\right]^{-1} \sum_{i=1}^{n} \mathbf{n}_i \left[\xi_{1_i}(\widehat{\boldsymbol{\theta}}^{(k)}) - \xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{(k+1)}\right] \\
&= \left(\mathbf{N}^\top \boldsymbol{\Omega}^{(k)} \mathbf{N} + \widehat{\alpha}^{(k)}\widehat{\sigma^2}^{(k)}\mathbf{K}\right)^{-1} \mathbf{N}^\top \left(\boldsymbol{\xi}_1^{(k)} - \boldsymbol{\Omega}^{(k)} \mathbf{X} \widehat{\boldsymbol{\beta}}^{(k+1)}\right), \\
\widehat{\sigma^2}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} \left[\xi_{2_i}(\widehat{\boldsymbol{\theta}}^{(k)}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}}^{(k)})\widehat{\mu}_i^{(k+1)} + \xi_{0_i}(\widehat{\boldsymbol{\theta}}^{(k)})\widehat{\mu_i}^{(k+1)^2}\right] \\
&= \frac{1}{n}\left(\mathbf{1}_n^\top \boldsymbol{\xi}_2^{(k)} - 2\widehat{\boldsymbol{\mu}}^{(k+1)\top} \boldsymbol{\xi}_1^{(k)} + \widehat{\boldsymbol{\mu}}^{(k+1)\top} \boldsymbol{\Omega}^{(k)} \widehat{\boldsymbol{\mu}}^{(k+1)}\right),
\end{aligned}
$$

3. ***CML-step***: Update $\boldsymbol{\nu}^{(k)}$ by maximizing the actual marginal log-likelihood function, obtaining

$$
\boldsymbol{\nu}^{(k+1)} = \underset{\boldsymbol{\nu}}{argmax} \left\{ \sum_{i=1}^{m} \log\left[F_{SMN}\left(\frac{\tau_i - \widehat{\mu}_i^{(k+1)}}{\widehat{\sigma}^{(k+1)}}\right)\right] + \sum_{i=m+1}^{n} \log\left[f_{SMN}(y_i|\widehat{\mu}_i^{(k+1)}, \widehat{\sigma^2}^{(k+1)}, \boldsymbol{\nu})\right] \right\}. \tag{2.1.14}
$$

The vector of parameters $\boldsymbol{\nu}$ is just a scalar (the degrees of freedom) for the Student-t and slash cases, while $\boldsymbol{\nu}^\top = (\varphi, \gamma)^\top$ for the contaminated normal case. A more efficient *CML-step* (2.1.14) can be easily accomplished by using, for instance, the `optimize` or `optimx` routines in the R software (R Core Team, 2017). The algorithm iterates between the *E*- and *CML-steps* until reaching convergence, i.e., until some distance involving two successive evaluations of the actual log-likelihood, like $||\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})||$ or $||\ell(\boldsymbol{\theta}^{(k+1)})/\ell(\boldsymbol{\theta}^{(k)}) - 1||$, is small enough. A set of reasonable starting values can be obtained by computing $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\sigma^2}^{(0)}$ as the solution of the least squares regression model of $\mathbf{Y}$ on $\mathbf{X}$, considering the censoring values as observed and $\widehat{\mathbf{f}}^{(0)} = (\mathbf{N}^\top\mathbf{N} + \alpha\widehat{\sigma^2}^{(0)}\mathbf{K})^{-1}\mathbf{N}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(0)})$.

2.1.3 Model selection and Estimation of α

In additive models, the Akaike information criterion (AIC) can be applied to select an appropriate α. Following Ferreira and Paula (2017), the AIC for PLR models is defined by:

$$
\text{AIC}(\alpha) = -2\ell_{cp}(\widehat{\boldsymbol{\theta}}, \alpha) + 2[p + q + \text{df}(\alpha)],
$$

where p is the dimension of the regression parameters $\boldsymbol{\beta}$, q the number of parameters of the SMN distribution being considered and $\ell_{cp}(\widehat{\boldsymbol{\theta}}, \alpha)$ is evaluated at $\widehat{\boldsymbol{\theta}}$ for a fixed α. The degrees of freedom (df) is defined as the number of effective parameters involved in modeling the nonparametric effects and can be approximated by (Hastie and Tibshirani, 1990):

$$
\text{df}(\alpha) = tr\{\mathbf{I}_r + \alpha Ł\},
$$

where $Ł = \widehat{\sigma^2}\mathbf{B}^{-1/2}\mathbf{K}\mathbf{B}^{-1/2}$, with $\mathbf{B} = \mathbf{N}^\top\mathbf{N}$.

2.1.4 Standard error approximation

In this subsection, we obtain the standard error approximation of the MPL estimates. Analogously to the parametric case, the approximate variance-covariance matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{f}^\top, \sigma^2)^\top$ is derived from the inverse of the observed information matrix (Mark et al., 1994). In effect, $\widehat{Var}_{approx}(\widehat{\boldsymbol{\theta}}) = \mathrm{I}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}|y)|_{\widehat{\boldsymbol{\theta}}}$, where $\mathrm{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}|y) = -\sum_{i=1}^{n}\frac{\partial^2\ell_{cp_i}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ and $\ell_{cp_i}(\boldsymbol{\theta})$ is the penalized log-likelihood function of the SMN-PCR model, given by:

$$\ell_{cp}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\ell_{cp_i}(\boldsymbol{\theta}) = \sum_{i=1}^{m}\log\left[\Psi_i(\boldsymbol{\theta})\right] + \sum_{i=m+1}^{n}\left\{-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 + \log[\psi_i(\boldsymbol{\theta})]\right\} - \frac{\alpha}{2}\mathbf{f}^\top\mathbf{K}\mathbf{f},$$

with

$$\Psi_i(\boldsymbol{\theta}) = \int_0^\infty \Phi\left[k^{-1/2}(u_i)D_i\right]dH(u_i|\boldsymbol{\nu}) \quad \text{and} \quad \psi_i(\boldsymbol{\theta}) = \int_0^\infty k^{-1/2}(u_i)\exp\left[\frac{-k^{-1}(u_i)d_i}{2}\right]dH(u_i|\boldsymbol{\nu}),$$

where $d_i = \frac{(y_i-\mu_i)^2}{\sigma^2}$ and $D_i = \sqrt{d_i}$. Thus, the matrix of second derivatives $\mathrm{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}|y)$ can be represented as:

$$\mathrm{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}|y) = -\sum_{i=1}^{n}\frac{\partial^2\ell_{cp_i}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = \mathrm{I}^1(\boldsymbol{\theta}) + \mathrm{I}^2(\boldsymbol{\theta}) + \mathrm{I}^3(\boldsymbol{\theta}),$$

where

$$\begin{aligned}
\mathrm{I}^1(\boldsymbol{\theta}) &= -\sum_{i=1}^{m}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log[\Psi_i(\boldsymbol{\theta})]\right\} = \sum_{i=1}^{m}\left[\frac{1}{\Psi_i^2(\boldsymbol{\theta})}\frac{\partial\Psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\Psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top} - \frac{1}{\Psi_i(\boldsymbol{\theta})}\frac{\partial^2\Psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right],\\
\mathrm{I}^2(\boldsymbol{\theta}) &= -\sum_{i=m+1}^{n}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{\alpha}{2(n-m)}\mathbf{f}^\top\mathbf{K}\mathbf{f}\right]\right\},\\
\mathrm{I}^3(\boldsymbol{\theta}) &= -\sum_{i=m+1}^{n}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log[\psi_i(\boldsymbol{\theta})]\right\} = \sum_{i=1}^{m}\left[\frac{1}{\psi_i^2(\boldsymbol{\theta})}\frac{\partial\psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top} - \frac{1}{\psi_i(\boldsymbol{\theta})}\frac{\partial^2\psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right].
\end{aligned}$$

The calculation of $\psi_i(\boldsymbol{\theta})$ and $\Psi_i(\boldsymbol{\theta})$ involves, respectively, the pdf and cdf of the normal, Student-t, slash and contaminated normal distributions, thus

$$\frac{\partial\Psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbb{I}_i^\Phi(1/2)\frac{\partial D_i}{\partial\boldsymbol{\theta}}, \qquad \frac{\partial^2\Psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = -\frac{1}{2}\mathbb{I}_i^\Phi(3/2)\frac{\partial D_i^2}{\partial\boldsymbol{\theta}}\frac{\partial D_i}{\partial\boldsymbol{\theta}} + \mathbb{I}_i^\Phi(1/2)\frac{\partial^2 D_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}, \tag{2.1.15}$$

$$\frac{\partial\psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = -\frac{1}{2}\mathbb{I}_i^\phi(3/2)\frac{\partial d_i}{\partial\boldsymbol{\theta}}, \qquad \frac{\partial^2\psi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = \frac{1}{4}\mathbb{I}_i^\phi(5/2)\frac{\partial d_i}{\partial\boldsymbol{\theta}}\frac{\partial d_i}{\partial\boldsymbol{\theta}^\top} - \frac{1}{2}\mathbb{I}_i^\phi(3/2)\frac{\partial^2 d_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}. \tag{2.1.16}$$

Using the same notation as in Lachos et al. (2011), we have that:

$$\mathbb{I}_i^{\phi}(\omega) = \int_0^{\infty} k^{-\omega}(u_i)\text{exp}\left[\frac{-k^{-1}(u_i)d_i}{2}\right] dH(u_i|\boldsymbol{\nu}) \tag{2.1.17}$$

Since $\mathbb{I}_i^{\Phi}(\omega) = \frac{1}{\sqrt{2\pi}}\mathbb{I}_i^{\phi}(\omega)$, for each distribution considered, the integral defined in (2.1.17) can be written as:

- Student-t distribution
$$\mathbb{I}_i^{\phi}(\omega) = \frac{\nu^{\nu/2}2^{\omega}\Gamma(\omega+\frac{\nu}{2})}{\Gamma\left(\frac{\nu}{2}\right)(\nu+d_i)^{\omega+\nu/2}};$$

- Slash distribution
$$\mathbb{I}_i^{\phi}(\omega) = \nu\int_0^1 u_i^{\omega+\nu-1}\text{exp}\left(-\frac{u_i}{2}d_i\right)du_i;$$

- Contaminated normal distribution

$$\mathbb{I}_i^{\phi}(\omega) = \sqrt{2\pi}\left[\varphi\gamma^{\omega-1/2}\phi(\sqrt{d_i};0,1/\gamma) + (1-\varphi)\phi(\sqrt{d_i};0,1)\right].$$

Let $\mathbf{x}_i^* = (\mathbf{x}_i^\top, \mathbf{n}_i^\top)^\top$, $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \mathbf{f}^\top)^\top$, for the Equations (2.1.15) and (2.1.16), the first and second derivatives of D_i, d_i for $\boldsymbol{\theta} = (\boldsymbol{\eta}, \sigma^2)$, using the notation $\ddot{D}_{\boldsymbol{\theta}}(D_i) = \dfrac{\partial D_i}{\partial\boldsymbol{\theta}}$, $D_{\boldsymbol{\theta}}(d_i) = \dfrac{\partial d_i}{\partial\boldsymbol{\theta}}$, $\ddot{D}^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}(D_i) = \dfrac{\partial^2 D_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$ and $D^2_{\boldsymbol{\theta}\boldsymbol{\theta}^\top}(d_i) = \dfrac{\partial^2 d_i}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$, are given by:

$\ddot{D}_{\eta}(D_i) = -\frac{\mathbf{x}_i^*}{\sigma}$, $\quad \ddot{D}_{\sigma^2}(D_i) = \frac{1}{2(\sigma^2)^{3/2}}(y_i - \mathbf{x}_i^{*\top}\boldsymbol{\eta})$,

$\ddot{D}^2_{\eta\eta^\top}(D_i) = \mathbf{0}$, $\quad \ddot{D}^2_{\sigma^2\sigma^2}(D_i) = \frac{3}{4(\sigma^2)^{5/2}}(y_i - \mathbf{x}_i^{*\top}\boldsymbol{\eta})$,

$\ddot{D}^2_{\eta\sigma^2}(D_i) = \frac{\mathbf{x}_i^*}{(\sigma^2)^{3/2}}$,

$D_{\eta}(d_i) = -\frac{1}{\sigma^2}(\mathbf{x}_i^* y_i - \mathbf{x}_i^*\mathbf{x}_i^{*\top}\boldsymbol{\eta})$, $\quad D_{\sigma^2}(d_i) = -\frac{1}{(\sigma^2)^2}(y_i - \mathbf{x}_i^{*\top}\boldsymbol{\eta})^2$,

$D^2_{\eta\eta^\top}(d_i) = \frac{2}{\sigma^2}\mathbf{x}_i^*\mathbf{x}_i^{*\top}$, $\quad D^2_{\sigma^2\sigma^2}(d_i) = \frac{2}{(\sigma^2)^3}(y_i - \mathbf{x}_i^{*\top}\boldsymbol{\eta})^2$ and

$D^2_{\eta\sigma^2}(d_i) = \frac{2}{(\sigma^2)^2}(\mathbf{x}_i^* y_i - \mathbf{x}_i^*\mathbf{x}_i^{*\top}\boldsymbol{\eta})$.

For $\text{I}^2(\boldsymbol{\theta})$, it is straightforward to find that $\text{I}^2_{\boldsymbol{\eta}\boldsymbol{\eta}^\top} = \alpha\mathbf{K}^*$, $\text{I}^2_{\boldsymbol{\eta}\sigma^2} = \mathbf{0}$ and $\text{I}^2_{\sigma^2\sigma^2} = -\dfrac{n-m}{2(\sigma^2)^2}$, where $\mathbf{K}^*$ is a block diagonal matrix of dimension $(p+r)\times(p+r)$, given by:

$$\mathbf{K}^* = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{K} \end{bmatrix}.$$

2.2 Diagnostic analysis

After the estimation procedure, the next step is evaluation of model results to detect outlying and influential observations in addition to possible deviations in the model, because in some cases the character of the regression can be determined only by a few observations. For example, in the context of simple linear regression, it is well known that inferences based on ordinary least squares regression can be strongly influenced by only a few outlying observations in the data. In semiparametric regression models, the effect of estimates under the influence of some observations is not an exception, the analysis can be influenced by minor perturbations of the model (Choongrak et al., 2002; Zhu et al., 2003; Hadi, 2016). In these circumstances, Cook and Weisberg (1982) stated there are two alternatives to handle this situation. The first consists of the development of robust estimation methods that require few assumptions and the second is related to the development of diagnostic tools to detect possible influential observations. So, we have the case-deletion approach (Cook, 1977), a traditional method for identifying influential observations, and the local influence approach, with the aim of investigating the behavior of some influence measures when we introduce small perturbations in the data and then monitor their impact on the outcome of the analysis.

2.2.1 Case deletion

Case-deletion is a widely used approach that studies the effect on the final inferential results of dropping the i-th case from the dataset. Hereafter, any subscript $[-i]$ refers to the original dataset with the i-th case deleted. In general, we consider $y_{[-i]} = (y_1, y_2, .., y_{n-1})^\top$ as the complete dataset with the i-th observation deleted. The complete-data penalized log-likelihood calculated after eliminating the i-th observation is denoted by $\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z}_{[-i]})$, therefore let $\widehat{\boldsymbol{\theta}}_{[-i]} = (\widehat{\boldsymbol{\beta}}^\top_{[-i]}, \widehat{\mathbf{f}}^\top_{[-i]}, \widehat{\sigma^2}_{[-i]})^\top$ be the argument that maximizes the function $Q_{[-i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \mathrm{E}[\ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z}_{[-i]})|y_{obs[-i]}, \hat{\boldsymbol{\theta}}]$, where $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\mathbf{f}}^\top, \widehat{\sigma^2})^\top$ are the MPL estimates obtained through ECME algorithm for $\boldsymbol{\theta}$.

To measure the influence of i-th observation in the MPL estimates of $\boldsymbol{\theta}$, we compare the difference between $\widehat{\boldsymbol{\theta}}_{[-i]}$ and $\widehat{\boldsymbol{\theta}}$. If this difference is large, then the i-th case can be considered influential, so it will require special attention. Since $\boldsymbol{\theta}_{[-i]}$ must be performed considering each individual separately for $i = 1, \ldots, n$, the computational effort can be high for large sample sizes. To circumvent this, Zhu and Lee (2001) proposed the following one-step pseudo approximation:

$$\widehat{\boldsymbol{\theta}}^*_{[-i]} = \widehat{\boldsymbol{\theta}} + \{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\dot{Q}_{[-i]}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) \tag{2.2.1}$$

where $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \dfrac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$ and $\dot{Q}_{[-i]}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = \dfrac{\partial Q_{[-i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$,

are the Hessian matrix and the individual score vector evaluated at $\widehat{\boldsymbol{\theta}}$, respectively.

Thus, $\dot{Q}_{[-i]}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = (\dot{Q}_{[-i]_{\boldsymbol{\beta}}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}),\ \dot{Q}_{[-i]_{\mathbf{f}}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}),\ \dot{Q}_{[-i]_{\sigma^2}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}))^\top$ has its elements as:

$$\begin{aligned}
\dot{Q}_{[-i]_{\boldsymbol{\beta}}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial Q_{[-i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \frac{1}{\widehat{\sigma^2}}\sum_{i\neq j}\Big[\xi_{1_j}(\widehat{\boldsymbol{\theta}})\mathbf{x}_j - \xi_{0_j}(\widehat{\boldsymbol{\theta}})\mathbf{x}_j\widehat{\mu}_j\Big],\\
\dot{Q}_{[-i]_{\mathbf{f}}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial Q_{[-i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \mathbf{f}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \frac{1}{\widehat{\sigma^2}}\sum_{i\neq j}\Big[\xi_{1_j}(\widehat{\boldsymbol{\theta}})\mathbf{n}_j - \xi_{0_j}(\widehat{\boldsymbol{\theta}})\mathbf{n}_j\widehat{\mu}_j\Big] - \frac{\widehat{\alpha}}{n}\mathbf{K}\widehat{\mathbf{f}},\\
\dot{Q}_{[-i]_{\sigma^2}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial Q_{[-i]}(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \sigma^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{2\widehat{\sigma^2}}\sum_{i\neq j}\Big\{1 - \frac{1}{\widehat{\sigma^2}}\Big[\xi_{2_j}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_j}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_j + \xi_{0_j}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_j^2\Big]\Big\}.
\end{aligned}$$

Following Zhu and Lee (2001) to measure the distance between $\widehat{\boldsymbol{\theta}}_{[-i]}$ and $\widehat{\boldsymbol{\theta}}$ and therefore to assess influential observations, we compute the *generalized Cook's distance* as follows:

$$GD_i = \Big(\widehat{\boldsymbol{\theta}}_{[-i]} - \widehat{\boldsymbol{\theta}}\Big)^\top\Big\{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\Big\}\Big(\widehat{\boldsymbol{\theta}}_{[-i]} - \widehat{\boldsymbol{\theta}}\Big), \qquad i = 1,\ldots,n \tag{2.2.2}$$

and by substituting Equation (2.2.2) into (2.2.1), we obtain the approximation of the *generalized Cook's distance*

$$GD_i^1 = \dot{Q}_{[-i]}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})^\top\Big\{-\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\Big\}^{-1}\dot{Q}_{[-i]}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) \qquad i = 1,\ldots,n. \tag{2.2.3}$$

Another measure used to calculate the difference between $\widehat{\boldsymbol{\theta}}_{[-i]}$ and $\widehat{\boldsymbol{\theta}}$, similar to the likelihood displacement defined in Cook (1986), is the Q-displacement (Zhu and Lee, 2001), defined as:

$$QD_i = 2\Big\{Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q(\widehat{\boldsymbol{\theta}}_{[-i]}|\widehat{\boldsymbol{\theta}})\Big\} \qquad i = 1,\ldots,n. \tag{2.2.4}$$

The Hessian matrix, $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$

After some rearrangement of terms and evaluation of the derivatives at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, we obtain the Hessian matrix $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ with elements given by:

$$\begin{aligned}
\ddot{Q}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^n \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{x}_i^\top,\\
\ddot{Q}_{\mathbf{f}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\mathbf{f}\partial\mathbf{f}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^n \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{n}_i\mathbf{n}_i^\top - \widehat{\alpha}\mathbf{K},\\
\ddot{Q}_{\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\sigma^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{n}{2\widehat{\sigma^2}} + \frac{1}{(\widehat{\sigma^2})^3}\sum_{i=1}^n\Big[\xi_{2_i}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^2\Big],\\
\ddot{Q}_{\boldsymbol{\beta}\mathbf{f}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\mathbf{f}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{\widehat{\sigma^2}}\sum_{i=1}^n \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\mathbf{n}_i^\top,\\
\ddot{Q}_{\boldsymbol{\beta}\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\sigma^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{(\widehat{\sigma^2})^2}\sum_{i=1}^n\Big[\xi_{1_i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i\widehat{\mu}_i\Big],\\
\ddot{Q}_{\mathbf{f}\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) &= \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\mathbf{f}\partial\sigma^2}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{1}{(\widehat{\sigma^2})^2}\sum_{i=1}^n\Big[\xi_{1_i}(\widehat{\boldsymbol{\theta}})\mathbf{n}_i - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{n}_i\widehat{\mu}_i\Big].
\end{aligned}$$

2.2.2 Local influence

Local influence analysis seeks to verify if small perturbations in the model or in the data affect the parameter estimates. Hence, to study the behavior of some influence measures, we will follow the approach proposed by Zhu and Lee (2001), where the Q-function is perturbed to assess the influence of this perturbation on the estimation. Ibacache-Pulgar and Paula (2011) and Ferreira and Paula (2017) applied this method successfully in the context of PLR models.

Consider a perturbation vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ restricted to some open subset $\boldsymbol{\Omega} \in \mathcal{R}^n$. Let $\ell_{cp}(\boldsymbol{\theta}, \boldsymbol{\omega}|\boldsymbol{z})$ be the complete-data penalized log-likelihood function of the perturbed model. Thus, we assume that a $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ exists such that $\ell_{cp}(\boldsymbol{\theta}, \boldsymbol{\omega}|\boldsymbol{z}) = \ell_{cp}(\boldsymbol{\theta}|\boldsymbol{z})$ for all $\boldsymbol{\theta}$. Also, let $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}) = (\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega})^\top, \widehat{\mathbf{f}}(\boldsymbol{\omega})^\top, \widehat{\sigma^2}(\boldsymbol{\omega}))^\top$ denote the maximum of the function $Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \mathrm{E}[\ell_{cp}(\boldsymbol{\theta}, \boldsymbol{\omega}|\boldsymbol{z})|\boldsymbol{Y}_{obs}, \widehat{\boldsymbol{\theta}}]$. Then, the influence graph is defined as $\boldsymbol{\alpha}(\boldsymbol{\omega}) = (\boldsymbol{\omega}^\top, f_Q(\boldsymbol{\omega}))^\top$, where $f_Q(\boldsymbol{\omega})$ is the Q-displacement function, defined as:

$$f_Q(\boldsymbol{\omega}) = 2\Big[Q(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) - Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})|\widehat{\boldsymbol{\theta}})\Big].$$

To approximate the Q-displacement function, the normal curvature $C_{f_Q,\boldsymbol{h}}(\boldsymbol{\theta})$ of $\boldsymbol{\alpha}(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$ in the direction of a unit vector $\boldsymbol{h}$ ($||\ \boldsymbol{h}\ || = 1$) is used to summarize the local behavior of $f_Q(\boldsymbol{\omega})$. It can be shown that:

$$C_{f_Q,\boldsymbol{h}}(\boldsymbol{\theta}) = -2\boldsymbol{h}^\top \ddot{Q}_{\boldsymbol{\omega}_0}\boldsymbol{h} \ = \ 2\boldsymbol{h}^\top \Delta^\top_{\boldsymbol{\theta},\boldsymbol{\omega}_0}\{\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}\boldsymbol{h},$$

leading to

$$-\ddot{Q}_{\boldsymbol{\omega}_0} = \Delta^\top_{\boldsymbol{\theta},\boldsymbol{\omega}_0}\{\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}.$$

Additionally, $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ is the Hessian matrix of dimension $(p + r + 1) \times (p + r + 1)$ and $\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0} = \partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^\top = (\Delta^\top_{\boldsymbol{\beta},\boldsymbol{\omega}_0}, \Delta^\top_{\mathbf{f},\boldsymbol{\omega}_0}, \Delta^\top_{\sigma^2,\boldsymbol{\omega}_0})^\top$ is the matrix of dimension $(p + r + 1) \times n$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

The information provided by $-\ddot{Q}_{\boldsymbol{\omega}_0}$ is fundamental for detecting influential observations (Cook, 1986). From the spectral decomposition of a symmetric matrix

$$-2\ddot{Q}_{\boldsymbol{\omega}_0} = \sum_{k=1}^{n} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top,$$

where $(\lambda_1, \mathbf{v}_1)$, $\cdots$, $(\lambda_n, \mathbf{v}_n)$ are the eigenvalue-eigenvector pairs of $-2\ddot{Q}_{\boldsymbol{\omega}_0}$ with $\lambda_1 \geq \ldots \geq \lambda_r$, $\lambda_{r+1} = \ldots = \lambda_n = 0$ and orthonormal eigenvectors $\{\mathbf{v}_k, k = 1, .., n\}$. Zhu and Lee (2001) and Lee and Xu (2004) proposed to examine all eigenvectors corresponding to nonzero eigenvalues to capture more information. For this end, we have the expressions

$$\tilde{\lambda}_k = \frac{\lambda_k}{\lambda_1 + \ldots + \lambda_r}, \ \ \mathbf{v}_k^2 = (v_{k1}^2, \ldots, v_{kn}^2)^\top \qquad \text{and} \qquad M(0) = \sum_{k=1}^{r} \lambda_k \mathbf{v}_k^2.$$

Let $M(0)_l = \sum_{k=1}^{r} \tilde{\lambda}_k v_{kl}^2$ denote the l-th component of $M(0)$. The evaluation of influential observations is based on the visual inspection of $M(0)_l$ plotted against the l-th index, for $l = 1, \ldots, n$. There are some disadvantages in using the normal curvature for influence analysis, since this measure may assume any value (not bounded), meaning it is not invariant under uniform scaling changes. Instead, we use the conformal normal curvature (Poon and Poon, 1999), given by:

$$B_{f_Q,\boldsymbol{h}}(\boldsymbol{\theta}) = \frac{C_{f_Q,\boldsymbol{h}}(\boldsymbol{\theta})}{\text{tr}\Big[-2\ddot{Q}_{\boldsymbol{\omega}_0}\Big]} \qquad \Rightarrow \qquad B_{f_Q,\boldsymbol{h}_l}(\boldsymbol{\theta}) = \frac{\Delta_{l_{\boldsymbol{\theta},\boldsymbol{\omega}_0}}^{\top}\{\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\Delta_{l_{\boldsymbol{\theta},\boldsymbol{\omega}_0}}}{\text{tr}\Big[\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}^{\top}\{\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})\}^{-1}\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}\Big]}, \tag{2.2.5}$$

where $\boldsymbol{h}_l$ is a column vector in $\mathcal{R}^n$ with the l-th entry equal to one, and zeros in the remaining positions. Here, $\Delta_{l_{\boldsymbol{\theta},\boldsymbol{\omega}_0}}^{\top}$ corresponds to $\boldsymbol{h}_l^{\top}\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}^{\top}$. The conformal normal curvature defined in (2.2.5) has the property that $0 \leq B_{f_Q,\boldsymbol{h}_l}(\boldsymbol{\theta}) \leq 1$ and the calculation is computationally easier. Based on the work of Zhu and Lee (2001), $M(0)_l$ can be obtained via $B_{f_Q,\boldsymbol{h}_l}(\boldsymbol{\theta})$ for all l.

Currently, there is no general rule for determining a benchmark value to indicate whether an observation is influential or not. Let $\overline{M(0)}$ and $SM(0)$ be the mean and standard error of $\{M(0)_l,\, l = 1, \ldots, n\}$ respectively. Zhu and Lee (2001) showed that $\overline{M(0)} = 1/n$ and proposed $\overline{M(0)} + 2SM(0)$ as benchmark value for $M(0)_l$. On the other hand, Lee and Xu (2004) presented a generalization and also proposed to use $M(0)_l > \overline{M(0)} + c^* SM(0)$, with c^* being a selected constant greater than 2. The choice of c^* depends on the context of the application (Ferreira and Paula, 2017), where the l-th case is considered as influential if $M(0)_l$ is larger than the benchmark.

Perturbation schemes

We now evaluate the matrix $\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}$ under four different perturbation schemes for the SMN-PCR model.

(a) Case-weight perturbation

This case of perturbation is appropriate to detect observations with large contribution to the penalized log-likelihood function and that may exercise strong influence on the maximum penalized likelihood estimates. From Equation (2.1.8), the so-called perturbed Q-function, considering an arbitrary attribution of weights, we have:

$$Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \omega_i \text{E}[\ell_{cp_i}(\boldsymbol{\theta}|\boldsymbol{z})|\boldsymbol{Y}_{obs}, \widehat{\boldsymbol{\theta}}] = \sum_{i=1}^{n} \omega_i Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) - \frac{\widehat{\alpha}}{2}\widehat{\mathbf{f}}^{\top}\mathbf{K}\widehat{\mathbf{f}}, \tag{2.2.6}$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)^{\top} \in \mathcal{R}^n$. The original expected value of the penalized complete-data log-likelihood corresponds to $\boldsymbol{\omega}_0 = (1, \ldots, 1)^{\top}$. In this perturbation scheme, the matrix $\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}$ derived

from (2.2.6) has elements given by:

$$\Delta_{\boldsymbol{\beta},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathbf{x}_i}{\widehat{\sigma^2}}\left[\xi_{1_i}(\widehat{\boldsymbol{\theta}}) - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i\right],$$

$$\Delta_{\mathbf{f},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\mathbf{f}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathbf{n}_i}{\widehat{\sigma^2}}\left[\xi_{1_i}(\widehat{\boldsymbol{\theta}}) - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i\right] \quad \text{and}$$

$$\Delta_{\sigma^2,\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = -\frac{1}{2\widehat{\sigma^2}} + \frac{1}{2(\widehat{\sigma^2})^2}\left[\xi_{2_i}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^2\right].$$

(b) Scale perturbation

In order to study the behavior of the estimates when there are possible deviations from the assumption of homogeneity, we assume that $Y_i \sim \text{SMN}(\mu_i, \sigma^2(\omega_i), \boldsymbol{\nu})$, with $\sigma^2(\omega_i) = \omega_i^{-1}\sigma^2$, $\omega_i > 0$ for $i = 1, \ldots, n$. The perturbed Q-function under this scheme is expressed as:

$$Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\left\{-\frac{1}{2}\log\left(\frac{\widehat{\sigma^2}}{\omega_i}\right) - \frac{\omega_i}{2\widehat{\sigma^2}}\left[\xi_{2_i}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^2\right]\right\} - \frac{\widehat{\alpha}}{2}\widehat{\mathbf{f}}^\top\mathbf{K}\widehat{\mathbf{f}}, \quad (2.2.7)$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)^\top \in \mathcal{R}^n$ is the vector of perturbations such that the non-perturbed model is obtained when $\boldsymbol{\omega}_0 = (1, \ldots, 1)^\top$. After some algebraic manipulations in (2.2.7), we obtain the elements of the matrix $\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}$ as:

$$\Delta_{\boldsymbol{\beta},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathbf{x}_i}{\widehat{\sigma^2}}\left[\xi_{1_i}(\widehat{\boldsymbol{\theta}}) - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i\right],$$

$$\Delta_{\mathbf{f},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\mathbf{f}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathbf{n}_i}{\widehat{\sigma^2}}\left[\xi_{1_i}(\widehat{\boldsymbol{\theta}}) - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i\right] \quad \text{and}$$

$$\Delta_{\sigma^2,\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{1}{2(\widehat{\sigma^2})^2}\left[\xi_{2_i}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^2\right].$$

(c) Explanatory variable perturbation

Here, we are interested in perturbing a specific continuous explanatory variable. Let $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)^\top$ be the vector of perturbations, $i = 1, \ldots, n$. So, the r-th explanatory variable of the design matrix is perturbed as $\mathbf{x}_{i\omega}^\top = \mathbf{x}_i^\top + \omega_i\mathcal{S}_r\mathbf{e}_r^\top$ for $r = 1, \ldots, p$, where $\mathcal{S}_r$ is the standard deviation of the r-th explanatory variable and $\mathbf{e}_r$ is a vector of dimension $p \times 1$, with a one in the r-th position and zero elsewhere. Under this scheme, we have that the perturbed Q-function is

$$Q(\boldsymbol{\theta}, \boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = -\frac{n}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^{n}\left[\xi_{2_i}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^* + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^{2*}\right] - \frac{\widehat{\alpha}}{2}\widehat{\mathbf{f}}^\top\mathbf{K}\widehat{\mathbf{f}}, \quad (2.2.8)$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)^\top \in \mathcal{R}^n$ and $\widehat{\mu}_i^* = \mathbf{x}_{i\omega}^\top\widehat{\boldsymbol{\beta}} + \mathbf{n}_i^\top\widehat{\mathbf{f}}$. Let $\boldsymbol{\omega}_0 = (0, \ldots, 0)^\top \in \mathcal{R}^n$ the vector of

non-perturbations. Taking the second derivative of (2.2.8) we find

$$\Delta_{\boldsymbol{\beta},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathcal{S}_r}{\widehat{\sigma^2}}\Big\{\xi_{1_i}(\widehat{\boldsymbol{\theta}})\mathbf{e}_r - \xi_{0_i}(\widehat{\boldsymbol{\theta}})\left[\widehat{\mu}_i\mathbf{e}_r + \mathbf{e}_r^\top\widehat{\boldsymbol{\beta}}\mathbf{x}_i\right]\Big\},$$

$$\Delta_{\mathbf{f},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\mathbf{f}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = -\frac{\mathcal{S}_r}{\widehat{\sigma^2}}\xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{e}_r^\top\widehat{\boldsymbol{\beta}}\ \mathbf{n}_i \quad \text{and}$$

$$\Delta_{\sigma^2,\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathcal{S}_r}{(\widehat{\sigma^2})^2}\left[\xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i\mathbf{e}_r^\top\widehat{\boldsymbol{\beta}} - \xi_{1_i}(\widehat{\boldsymbol{\theta}})\mathbf{e}_r^\top\widehat{\boldsymbol{\beta}}\right].$$

(d) Response variable perturbation

In this case, to perturb the response variable values we replace Y_{obs_i} by $Y_{\text{obs}_i}(\omega_i) = Y_{\text{obs}_i} + \mathcal{S}_y\omega_i$ for $i = 1,\ldots,n$, where $\mathcal{S}_y$ is the standard deviation of Y_{obs_i}. For the SMN-PCR model presented in Equations (2.1.1) and (2.1.2) we have:

$$Y_{\text{obs}_i}(\omega_i) = \begin{cases} \tau_i(\omega_i) & \text{if} \quad Y_i \leq \tau_i; \\ Y_i(\omega_i) & \text{if} \quad Y_i > \tau_i. \end{cases}$$

Therefore, $Y_i(\omega_i) = Y_i - \mathcal{S}_y\omega_i$ (Matos et al., 2013). So, the perturbed Q-function it is obtained replacing Y_{obs_i} values by $Y_{\text{obs}_i}(\omega_i)$, where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_n)^\top \in \mathcal{R}^n$ denotes the vector of perturbations and $\boldsymbol{\omega}_0 = (0,\ldots,0)^\top$ is the corresponding non-perturbation vector such that $Q(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) = Q(\boldsymbol{\theta},|\widehat{\boldsymbol{\theta}})$, with

$$\begin{aligned} Q(\boldsymbol{\theta},\boldsymbol{\omega}|\widehat{\boldsymbol{\theta}}) &= -\frac{n}{2}\log(\widehat{\sigma^2}) - \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^{n}\Big[\xi_{2_i}(\widehat{\boldsymbol{\theta}}) - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\mathcal{S}_y\omega_i + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathcal{S}_y^2\omega_i^2 - 2\xi_{1_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i \\ &+ 2\xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathcal{S}_y\omega_i\widehat{\mu}_i + \xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i^2\Big] - \frac{\widehat{\alpha}}{2}\widehat{\mathbf{f}}^\top\mathbf{K}\widehat{\mathbf{f}}. \end{aligned} \tag{2.2.9}$$

The matrix $\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}$ has elements given by:

$$\Delta_{\boldsymbol{\beta},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\beta}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = -\frac{\mathcal{S}_y}{\widehat{\sigma^2}}\ \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{x}_i,$$

$$\Delta_{\mathbf{f},\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\mathbf{f}\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = -\frac{\mathcal{S}_y}{\widehat{\sigma^2}}\ \xi_{0_i}(\widehat{\boldsymbol{\theta}})\mathbf{n}_i \quad \text{and}$$

$$\Delta_{\sigma^2,\boldsymbol{\omega}_0} = \frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\sigma^2\partial\omega_i}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = \frac{\mathcal{S}_y}{(\widehat{\sigma^2})^2}\ \left[\xi_{0_i}(\widehat{\boldsymbol{\theta}})\widehat{\mu}_i - \xi_{1_i}(\widehat{\boldsymbol{\theta}})\right].$$

Local influence in sub-vectors

Cook (1986) extended the analysis of local influence to a subset of the parameters of interest. Let us consider the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top,\boldsymbol{\theta}_2^\top)^\top$, where $\boldsymbol{\theta}_1 = \boldsymbol{\beta}$ and $\boldsymbol{\theta}_2 = (\mathbf{f}^\top,\sigma^2)^\top$. In PLR models for complete data, some authors, such as Zhu et al. (2003), Ibacache-Pulgar and Paula (2011), Chen et al. (2012) and Relvas and Paula (2016) have carried out local influence analysis on sub-vectors. In the context of generalized linear mixed models with missing data, Zhu and Lee (2001) demonstrated that $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) \approx \ddot{L}_{cp}$, where $\ddot{L}_{cp}$ is the penalized Hessian matrix. Thus, we can define an approximation to the

partial conformal curvature of $\boldsymbol{\theta}_1 = \boldsymbol{\beta}$ in the unitary direction $\boldsymbol{h}$ as:

$$C_{f_Q,\boldsymbol{h}}(\boldsymbol{\beta}) = \boldsymbol{h}^\top \Delta^\top_{\boldsymbol{\theta},\boldsymbol{\omega}_0} \left\{ [\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})]^{-1} - \ddot{Q}^{\widehat{\theta_2}} \right\} \Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0} \boldsymbol{h},$$

where $\ddot{Q}^{\widehat{\theta_2}} = \text{blockdiag}\{\mathbf{0}, (\ddot{Q}^{\widehat{\mathbf{f}}})^{-1}, (\ddot{Q}^{\widehat{\sigma^2}})^{-1}\}$ obtained from the partition of $\ddot{Q}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ according to $\boldsymbol{\theta}$, so $\ddot{Q}^{\widehat{\mathbf{f}}} = \ddot{Q}_{\mathbf{f}}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$ and $\ddot{Q}^{\widehat{\sigma^2}} = \ddot{Q}_{\sigma^2}(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}})$. Therefore, it is possible to define an approximation to the partial conformal normal curvature $B_{f_Q,\boldsymbol{h}_l}(\boldsymbol{\beta})$ based on Equation (2.2.5). The details of the proof are omitted.

Chapter 3

Results

In order to examine the performance of our proposed models and algorithm, in this chapter we present some simulation studies through a Monte Carlo (MC) experiment and analyze a real dataset. In relation to simulation studies, the first study is related to the parameter recovery and robustness of the MPL estimates. The second simulation study evaluate the finite-sample performance of the parameter estimates, e.i., we investigate the asymptotic properties of the MPL estimates from different cases of the SMN-PCR model. The last study shows the capacity of the diagnostic measures in detect potentially influential observations. Finally, we analyze a real dataset, the wage rate (Mroz, 1987).

3.1 Simulation study

In this section, the performance of the proposed algorithm is evaluated via simulation studies. These computational procedures were implemented using the `R` software (R Core Team, 2017). In particular, we consider the following PR model

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1.1}$$

where $\varepsilon_i \overset{\text{iid.}}{\sim} \text{SMN}(0, \sigma^2, \boldsymbol{\nu})$. We generated left-censored samples from the model given in (3.1.1) considering censoring levels 0%, 10%, 20% and 30% and sample sizes $n = 200, 300, 400$ and 600. For each combination of censoring level and sample size, we generated 500 samples from the SMN-PCR model, in four different situations: N-PCR, T-PCR ($\nu = 4$), SL-PCR ($\nu = 2$) and CN-PCR ($\boldsymbol{\nu}^\top = (0.1, 0.1)^\top$), for normal, Student-t, slash and contaminated normal distribution, respectively.

We performed all MC simulations setting $\boldsymbol{\beta} = (2, 4)^\top$ and $\sigma^2 = 2$, with $\mathbf{x}_i = (x_{1_i}, x_{2_i})^\top$ generated independently from uniform distributions on the intervals $U(0, 1)$ and $U(1, 2)$, respectively. The true nonparametric function was chosen as $f(t) = 10\,\text{sin}(2\pi t)$, with $t \in (0, 1.5)$. We also assumed different values for t, so the incidence matrix $\mathbf{N}$ was the identity matrix of order $n \times n$.

3.1.1 Parameter recovery and robustness of the MPL estimates.

We computed for each element of the parameter vector of interest $\boldsymbol{\theta}_k = (\beta_1, \beta_2, \sigma^2)^\top$, its MC mean $(\overline{\widehat{\theta}}_k)$ and MC standard deviation (MC-SD) for its MPL estimates. Also, the average values of the approximate standard errors obtained following Subsection 2.1.4 (OM-SD) were recorded for comparison purposes. So, we have that:

$$\overline{\widehat{\theta}}_k = \frac{1}{500}\sum_{j=1}^{500}\widehat{\theta}_k^{(j)} \qquad \text{MC-SD} = \sqrt{\sum_{j=1}^{500}(\widehat{\theta}_k^{(j)} - \overline{\widehat{\theta}}_k)^2/499} \qquad \text{OM-SD} = \frac{1}{500}\sum_{j=1}^{500}\text{SE}(\widehat{\theta}_k^{(j)})$$

Table 3.1 presents the MPL estimates of θ_k in the different scenarios considered. This table shows that the model fits provide estimates that are close to the true values of the parameters and are less sensitive to the variation of the censoring level. Besides this, the empirical standard deviations (MC-SD) are close to the MC standard errors (OM-SD) and the difference tends to zero as the sample size increases, indicating that the result of the standard errors (Subsection 2.1.4) is reliable. The results of the coverage probability are presented in Table 3.2, where 95% confidence intervals were computed for each scenario using the OM-SD. As expected, the "coverage probability" is stable and around 90% for the regression parameters β_1 and β_2, but the percentage of coverage is somewhat impaired for the intervals built for σ^2. Overall, the ECME algorithm produces satisfactory estimates for the SMN-PCR fitted models.

For the estimates of the nonparametric component, Figure 3.1 presents the behavior of the 500 MC samples under the T-PCR model. We can note from this figure that when the censoring level increases, the variability among the nonparametric estimates functions increases, however as the sample size increases, the variability decreases. Similar results were obtained for the other models (see also the Figures A.1, A.2 and A.3 of Appendix A).

On the other hand, we evaluate the robust aspects of the MPL estimates in the SMN-PCR model in the presence of outliers in the response variable. From the simulation scenarios previously considered, we generate 100 MC samples of size $n = 200$ under N-PCR model with left censored values, $\alpha = 0.0001$ and two levels of censoring, 10% and 20%. The goal is study the influence of the change of η units in a single observation. In this case, we contaminate and replace the observation #66 with $y_{66}(\eta) = y_{66} + \eta$, for $\eta \in \{2, 4, 6, 8, 10, 12, 14, 16\}$. In each replication, we obtained the MPL estimates with and without the contaminated values. We define the relative change of the parameter of interest as:

$$RC(\widehat{\theta}_i) = \left|\frac{\widehat{\theta}_i(\eta) - \widehat{\theta}_i}{\widehat{\theta}_i}\right|,$$

where $\widehat{\theta}_i(\eta)$ is the MPL estimate obtained with the contaminated dataset and $\widehat{\theta}_i$ is the estimate with the initial dataset. Figure 3.2 illustrates the mean values of the relative changes for estimates of β_1, β_2 and σ^2 for each value of η under N-PCR, T-PCR, SL-PCR and CN-PCR models respectively. As expected, we can see that for the all parameters, the mean values of relative changes under the N-PCR model increase significantly as the value of η increases, i.e. the estimates obtained in the N-PCR model are

Table 3.1: Simulated data. Mean value, MC standard deviation (MC-SD) and approximated standard errors (OM-SD) based in 500 artificial samples from the SMN-PCR model, considering left censoring.

		Model/Measure											
		N-PCR			**T-PCR**			**SL-PCR**			**CN-PCR**		
Parameter	C.L.	$\widehat{\bar{\theta}}_k$	MC-SD	OM-SD	$\widehat{\bar{\theta}}_k$	MC-SD	OM-SD	$\widehat{\bar{\theta}}_k$	MC-SD	OM-SD	$\widehat{\bar{\theta}}_k$	MC-SD	OM-SD
						n=200							
β_1	0%	1.986	0.373	0.339	2.060	0.448	0.405	2.008	0.459	0.449	2.045	0.421	0.381
	10%	1.983	0.384	0.347	2.070	0.446	0.412	2.010	0.459	0.456	2.035	0.434	0.390
	20%	1.990	0.410	0.367	2.080	0.487	0.438	2.013	0.490	0.483	2.055	0.470	0.414
	30%	2.008	0.429	0.393	2.080	0.506	0.468	2.028	0.535	0.516	2.082	0.493	0.441
β_2	0%	4.072	0.360	0.327	4.134	0.420	0.391	4.087	0.444	0.435	4.115	0.397	0.369
	10%	4.080	0.385	0.345	4.146	0.457	0.412	4.086	0.462	0.454	4.114	0.420	0.389
	20%	4.087	0.406	0.359	4.174	0.468	0.428	4.117	0.492	0.476	4.151	0.446	0.405
	30%	4.116	0.402	0.385	4.224	0.471	0.460	4.168	0.504	0.507	4.188	0.433	0.434
σ^2	0%	1.712	0.199	0.175	1.689	0.232	0.231	1.740	0.231	0.210	1.687	0.238	0.212
	10%	1.702	0.213	0.183	1.674	0.243	0.239	1.727	0.241	0.219	1.677	0.242	0.221
	20%	1.691	0.218	0.192	1.663	0.253	0.252	1.727	0.252	0.231	1.673	0.254	0.234
	30%	1.704	0.233	0.206	1.672	0.263	0.270	1.721	0.259	0.246	1.675	0.273	0.251
						n=300							
β_1	0%	1.985	0.290	0.279	2.048	0.355	0.333	2.000	0.353	0.369	2.032	0.335	0.315
	10%	1.980	0.291	0.286	2.052	0.360	0.341	2.001	0.365	0.378	2.033	0.339	0.323
	20%	1.983	0.304	0.302	2.054	0.387	0.362	2.016	0.390	0.397	2.037	0.354	0.342
	30%	1.998	0.325	0.323	2.084	0.402	0.385	2.041	0.412	0.425	2.063	0.374	0.365
β_2	0%	4.096	0.294	0.271	4.104	0.348	0.323	4.082	0.394	0.358	4.079	0.327	0.305
	10%	4.093	0.312	0.283	4.116	0.366	0.337	4.084	0.416	0.371	4.085	0.353	0.319
	20%	4.113	0.327	0.299	4.142	0.388	0.357	4.098	0.444	0.392	4.101	0.384	0.338
	30%	4.140	0.321	0.317	4.200	0.375	0.378	4.151	0.431	0.416	4.141	0.378	0.358
σ^2	0%	1.796	0.163	0.149	1.788	0.197	0.198	1.815	0.188	0.177	1.785	0.190	0.181
	10%	1.788	0.171	0.156	1.781	0.208	0.206	1.810	0.198	0.186	1.784	0.203	0.191
	20%	1.782	0.181	0.164	1.787	0.224	0.218	1.802	0.209	0.196	1.786	0.215	0.202
	30%	1.790	0.195	0.177	1.785	0.244	0.248	1.805	0.225	0.209	1.786	0.230	0.217
						n=400							
β_1	0%	1.998	0.247	0.238	2.039	0.298	0.284	2.011	0.305	0.315	2.033	0.275	0.268
	10%	1.997	0.253	0.245	2.041	0.298	0.291	2.008	0.309	0.323	2.026	0.280	0.275
	20%	2.004	0.268	0.261	2.066	0.327	0.311	2.022	0.336	0.344	2.050	0.298	0.294
	30%	2.012	0.292	0.280	2.072	0.365	0.333	2.053	0.361	0.368	2.065	0.318	0.315
β_2	0%	4.087	0.279	0.249	4.127	0.316	0.296	4.115	0.355	0.328	4.065	0.296	0.280
	10%	4.080	0.289	0.260	4.133	0.319	0.309	4.104	0.370	0.341	4.081	0.304	0.292
	20%	4.113	0.305	0.277	4.184	0.338	0.329	4.117	0.397	0.363	4.117	0.323	0.312
	30%	4.123	0.300	0.293	4.229	0.314	0.347	4.186	0.390	0.384	4.152	0.314	0.330
σ^2	0%	1.842	0.138	0.132	1.832	0.165	0.174	1.855	0.163	0.156	1.829	0.164	0.160
	10%	1.839	0.147	0.138	1.827	0.173	0.182	1.851	0.170	0.164	1.818	0.166	0.167
	20%	1.836	0.157	0.147	1.823	0.184	0.192	1.849	0.178	0.173	1.824	0.186	0.178
	30%	1.839	0.167	0.157	1.837	0.197	0.207	1.848	0.195	0.185	1.830	0.194	0.191
						n=600							
β_1	0%	1.996	0.206	0.196	2.027	0.239	0.232	2.021	0.253	0.258	2.012	0.214	0.220
	10%	2.001	0.214	0.204	2.040	0.238	0.241	2.018	0.262	0.266	2.006	0.226	0.230
	20%	2.004	0.226	0.217	2.061	0.264	0.259	2.030	0.277	0.284	2.028	0.244	0.244
	30%	2.010	0.244	0.232	2.062	0.287	0.275	2.053	0.295	0.304	2.027	0.261	0.261
β_2	0%	4.061	0.220	0.203	4.076	0.257	0.241	4.085	0.277	0.268	4.087	0.230	0.228
	10%	4.062	0.233	0.211	4.091	0.265	0.252	4.096	0.301	0.278	4.076	0.242	0.236
	20%	4.105	0.255	0.225	4.110	0.275	0.266	4.109	0.319	0.297	4.111	0.253	0.256
	30%	4.118	0.250	0.240	4.164	0.277	0.286	4.160	0.327	0.317	4.151	0.263	0.270
σ^2	0%	1.891	0.117	0.110	1.880	0.138	0.145	1.892	0.134	0.130	1.882	0.136	0.134
	10%	1.882	0.121	0.116	1.882	0.150	0.152	1.885	0.138	0.136	1.870	0.144	0.140
	20%	1.881	0.129	0.122	1.873	0.156	0.161	1.878	0.146	0.143	1.881	0.156	0.149
	30%	1.885	0.136	0.131	1.883	0.165	0.172	1.882	0.156	0.153	1.867	0.162	0.158

more sensitive to outliers. In addition, we may note that the estimates of the parameters β_1 and σ^2 are more affected by the presence of outliers and is greater when it increases the level of censoring (20%), while β_2 is more stable. In contrast, the SMN-PCR models with heavy tails (T, SL and CN) are less affected by variations of η and therefore more robust than those under N-PCR.

Table 3.2: Simulated data. Coverage probability $(\%)$ based on 500 samples from the SMN-PCR model, considering different left censoring levels (LCs).

		N-PCR			**T-PCR**			**SL-PCR**			**CN-PCR**		
n	LCs	β_1	β_2	σ^2	β_1	β_2	σ^2	β_1	β_2	σ^2	β_1	β_1	σ^2
200	0%	92.4	91.2	88.8	93.4	92.6	89.0	95.6	94.2	89.6	93.0	92.0	81.8
	10%	93.0	90.6	88.0	93.2	90.6	87.0	95.2	93.8	88.0	92.2	92.6	82.8
	20%	93.4	91.0	88.2	91.8	91.0	86.0	95.6	91.2	89.8	92.8	91.6	85.0
	30%	93.2	93.2	83.0	93.2	93.0	80.6	94.8	92.0	81.6	92.2	93.0	87.0
300	0%	94.2	92.2	88.8	93.2	93.0	86.6	97.2	92.2	85.8	93.2	93.0	84.8
	10%	93.8	91.4	90.2	94.0	92.0	88.4	96.8	92.0	86.8	94.6	92.0	84.6
	20%	95.2	91.4	81.2	93.2	91.6	89.2	95.8	91.4	88.0	95.2	90.0	85.0
	30%	96.2	93.2	84.0	93.0	92.6	88.4	96.6	92.2	87.0	94.6	92.8	87.6
400	0%	94.6	91.0	83.4	94.6	90.4	93.6	95.6	91.2	91.8	94.8	93.2	88.0
	10%	95.0	90.8	85.0	94.8	93.8	92.2	95.2	92.6	90.6	95.4	94.0	87.2
	20%	96.0	90.0	84.6	92.6	90.6	90.6	96.2	92.2	91.6	94.8	93.0	88.6
	30%	95.2	92.4	87.6	93.2	91.0	90.8	95.2	92.2	91.4	95.6	93.6	91.2
600	0%	94.6	92.6	92.4	95.2	91.8	90.5	96.2	94.0	93.6	96.0	95.2	92.2
	10%	95.8	92.2	90.2	96.6	94.4	95.6	96.2	92.8	92.2	95.8	93.4	91.8
	20%	93.6	91.0	88.8	94.8	93.4	95.8	96.6	92.8	95.2	94.6	94.4	92.4
	30%	92.6	91.8	91.4	95.2	92.8	90.4	95.6	92.0	94.6	94.2	93.6	94.8

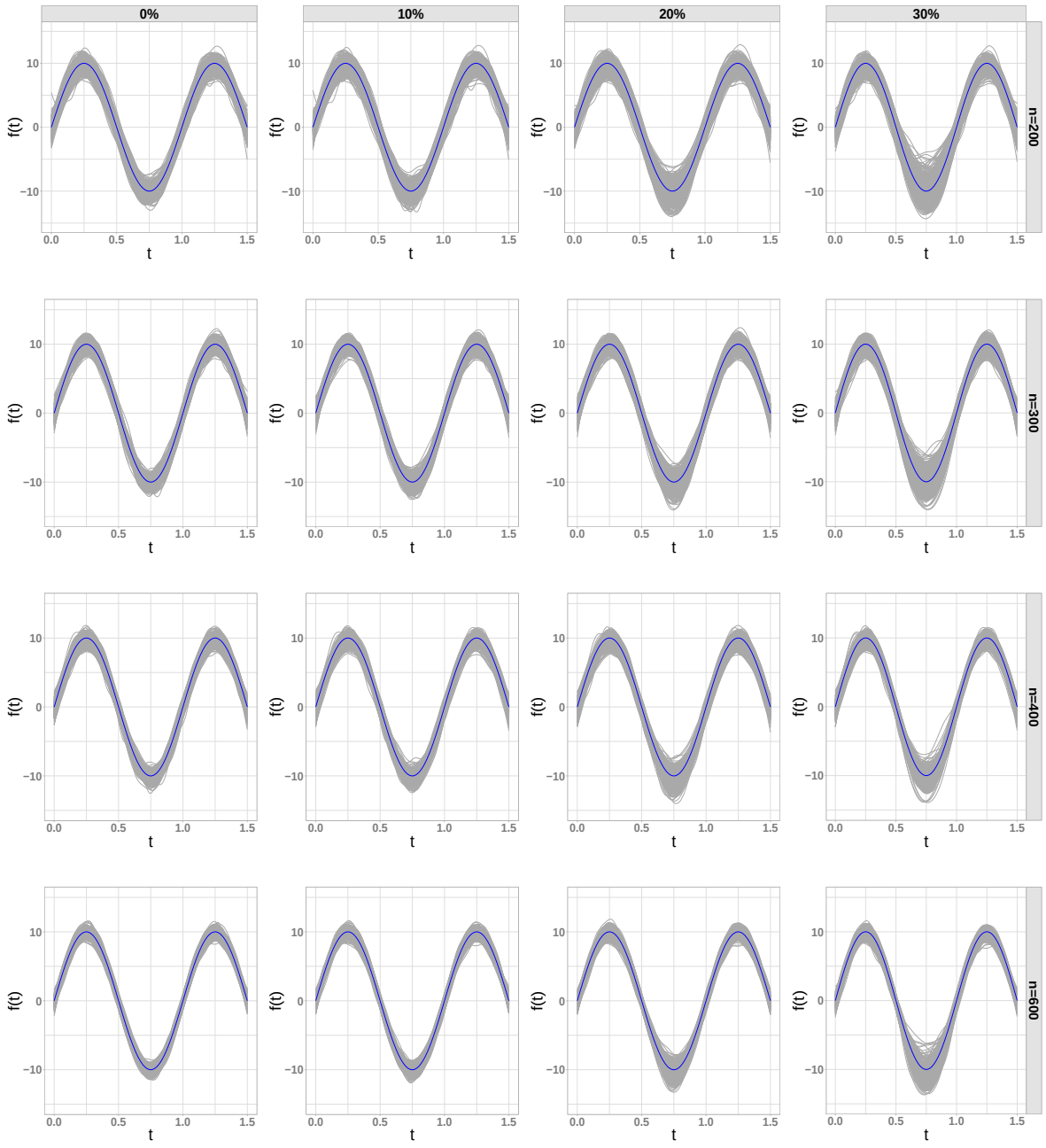

Figure 3.1: Simulated data. Behavior of the nonparametric component based on 500 samples from the T-PCR model. True curve (blue line) and adjusted curves (gray lines).

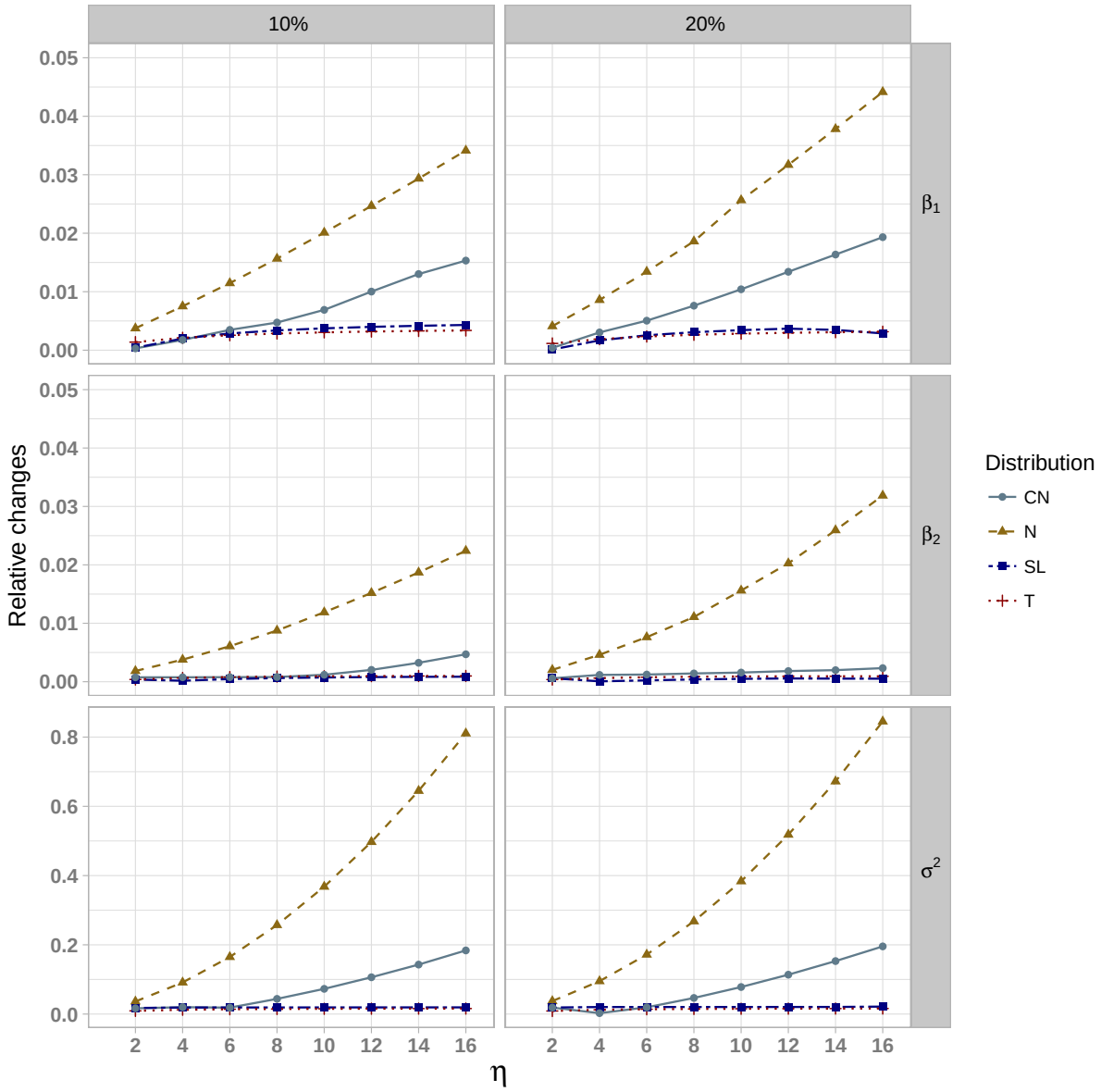

Figure 3.2: Simulated data. Mean values of the relative changes on the MPL estimates fitting a N-PCR, T-PCR, SL-PCR and CN-PCR models for different values of η on the observation 66.

3.1.2 Asymptotic properties

Now, to provide empirical evidence about the consistency of the MPL estimates. We analyzed the mean of the absolute bias (Bias) and mean of the Mean square error (MSE) of MPL estimated obtained from the fitted models in each Monte Carlo experiment. The Bias and MSE measures are given by:

$$\text{Bias}(\theta_k) = \frac{1}{500}\sum_{j=1}^{500}\left|\widehat{\theta}_k^{(j)} - \theta_k\right| \qquad \text{and} \qquad \text{MSE}(\theta_k) = \frac{1}{500}\sum_{j=1}^{500}(\widehat{\theta}_k^{(j)} - \theta_k)^2,$$

where $\widehat{\theta}_k^{(j)}$ is the MPL estimates of the parameter θ_k for the j-th MC sample, $j = 1, \ldots, 500$. Figures 3.3 and 3.4 shows a graphical representation of the asymptotic properties. From the scenarios considered,

the least appropriate was for $n = 200$ and we can see that the estimates obtained between the uncensored and most censored (30%) cases are quite similar. This gap is tolerably small and it becomes smaller when simple size increases. For instance, σ^2 estimates for a 600 sample size, are almost equal regardless of the censoring level and model. Thus, the results illustrate that Bias and MSE decrease as the sample size increases and we can see that, independently of level of censoring and type of SMN distribution, the properties for β_1, β_2 and σ^2 following patterns of convergence to zero. Hence, as a general rule the MPL estimates based on the proposed ECME algorithm for the SMN-PCR models, do show desirable asymptotic properties.

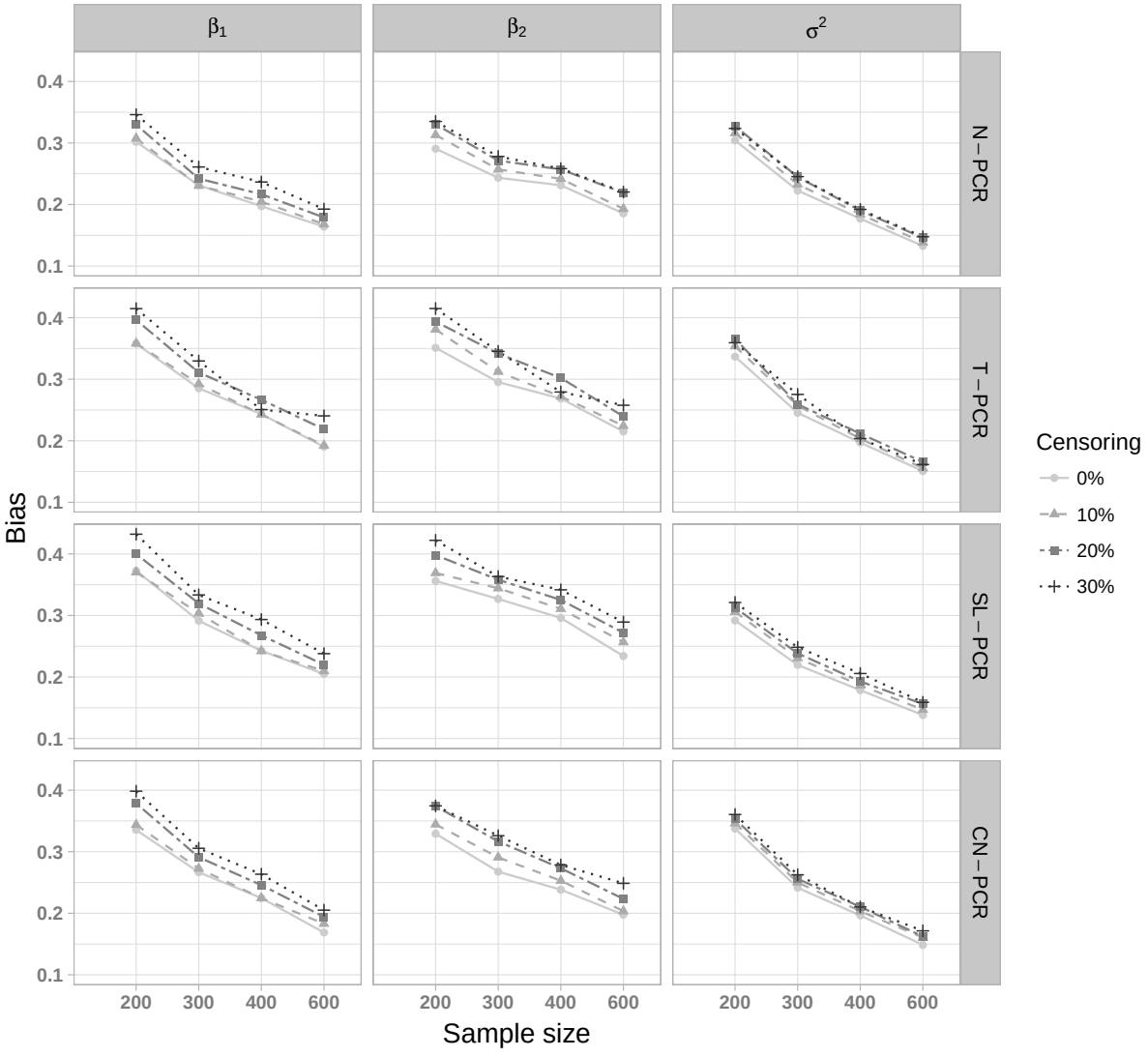

Figure 3.3: Simulated data. Asymptotic properties. MC mean of bias for β_1, β_2 and σ^2 for different sample sizes and levels of censoring in SMN-PCR models.

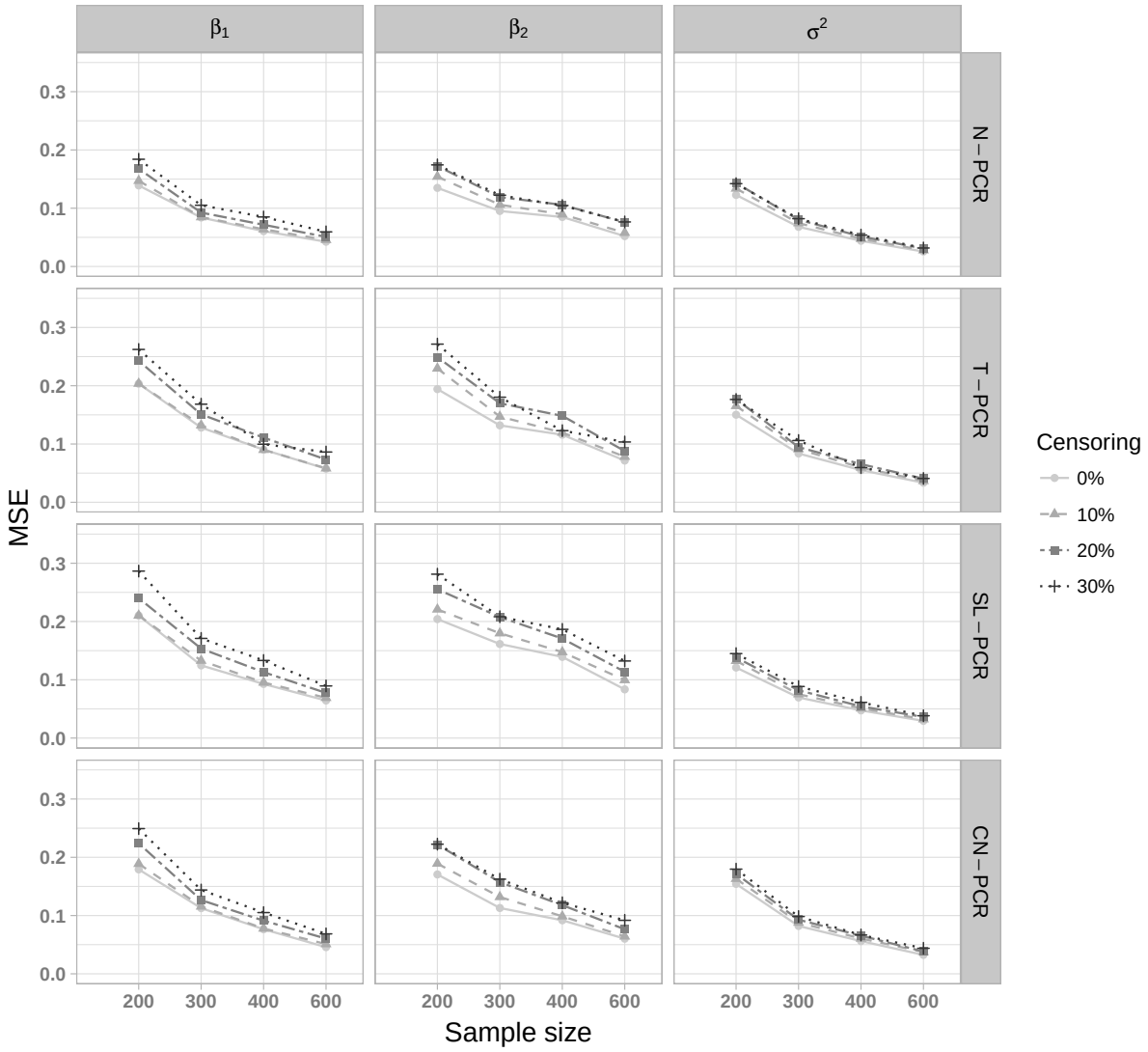

Figure 3.4: Simulated data. Asymptotic properties. MC mean of the Mean square error (MSE) for β_1, β_2 and σ^2 for different sample sizes and levels of censoring in SMN-PCR models.

3.1.3 Diagnostic measures

Additionally, we illustrate the capacity of the proposed diagnostic measures to identify possible influential observations. The diagnostic measures were computed from 200 MC simulations from a N-PCR model considering a sample size $n = 200$, 10% censoring level, same nonparametric function $f(t)$ as before and the benchmark setting at $c^* = 3.5$ (Subsection 2.2.2), for $i = 1, \ldots, 200$. Under this scenario, we contaminated observation 82 as follows

(i) Replace in the parametric component $\boldsymbol{\beta}$ by $2\boldsymbol{\beta}$ to generate the response of the observation #82 $\longrightarrow y_{82}$.

(ii) Replace $\boldsymbol{\beta}$ by $4\boldsymbol{\beta}$.

(iii) Replace $\boldsymbol{\beta}$ by $8\boldsymbol{\beta}$.

Table 3.3 presents the percentage of times that observation #82 was correctly identified as the most influential under different perturbation schemes considering the N-PCR model, and the percentage of times that a lower weight $u_{82} = \xi_{0_{82}}(\widehat{\boldsymbol{\theta}})$ was assigned with heavy-tailed models, such as, Student-t (T), slash (SL) and contaminated normal (CN). As expected, all percentages increase for higher contamination rates. First group (success percentages) represents the ability to identify influential observations in the normal model and the second one (preference percentages) indicates the robustness of heavy-tailed distributions, since a smaller weight is attributed to influential observations and it increases due to the high contamination rate. By way of illustration, Figures 3.5 and 3.6 present the behavior of the diagnostic measures through index plots for the case-weight and explanatory variable perturbation for one of the MC simulations (simulation No.3). Is clear the high sensitivity of the MPL estimates in the presence of influential observations when a N-PCR model is used, however are more robust under SMN-PCR models with heavy tails (T, SL and CN), since the observation #82 for these cases was not detected. Similar results are found for the other schemes (see Figures A.4 and A.5 of Appendix A)).

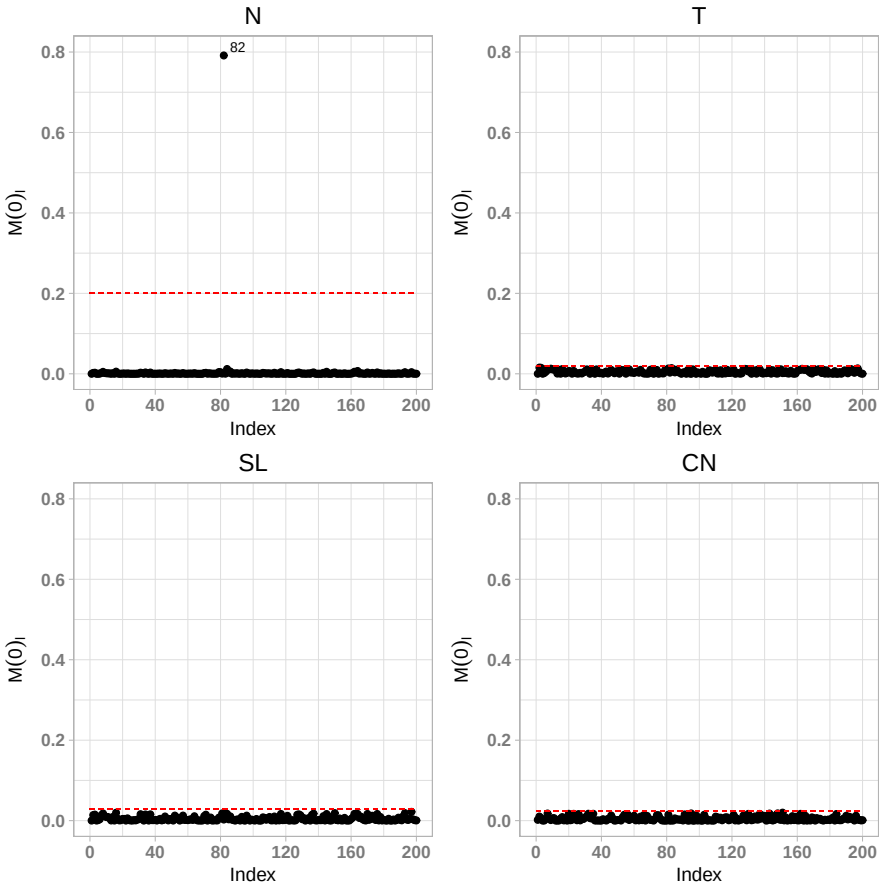

Figure 3.5: Simulated data. Index plot of $M(0)_l$ for assessing local influence for contamination rate 8β. Case-weight perturbation (simulation No.3).

Table 3.3: Simulated data. Success percentages for different perturbation schemes in the N-PCR model and preference percentages under the T, SL and CN models, for different contamination schemes.

	Normal				Heavy-tailed		
Contamination	Case-weight	Scale	Explanatory	Response	T	SL	CN
2β	71.5	71.5	72.5	70.5	71.5	70.5	71.5
4β	95.5	95.5	94.5	89.0	94.0	94.0	94.0
8β	97.5	97.5	96.5	96.5	96.5	96.5	95.5

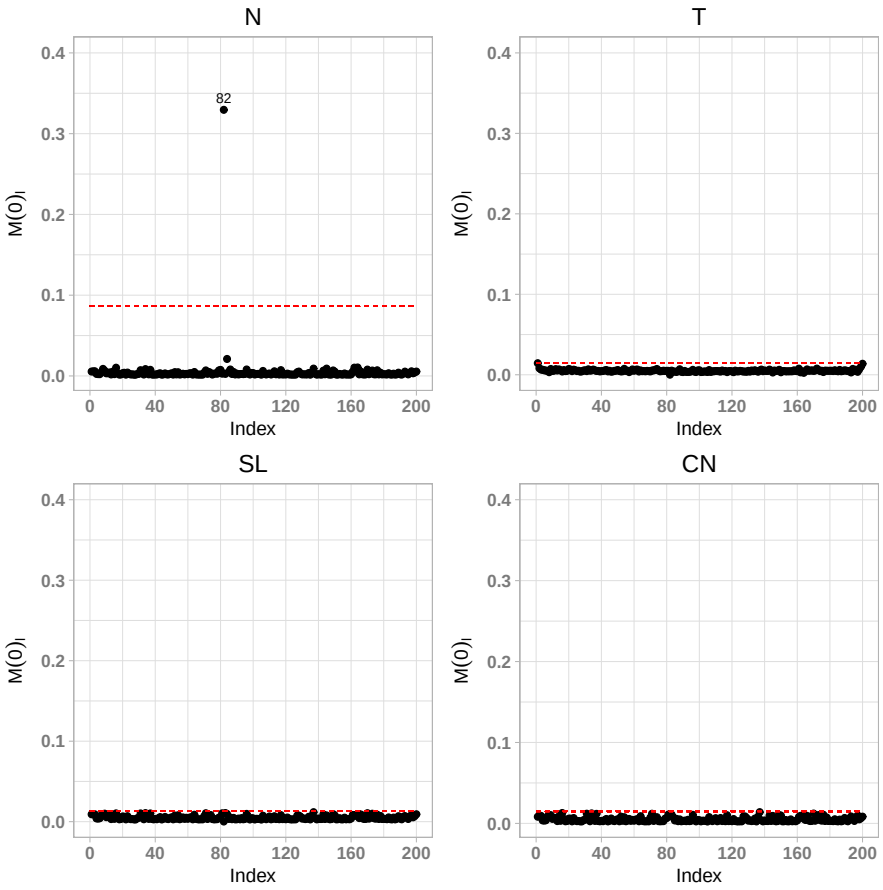

Figure 3.6: Simulated data. Index plot of $M(0)_l$ for assessing local influence for contamination rate 8β. Explanatory variable perturbation (simulation No.3).

3.2 Application: Wage rate data

In this section, we illustrate the performance of the proposed method by analyzing the wage rate dataset described in Mroz (1987). The dataset comes from the University of Michigan Panel Study of Income Dynamics (PSID) and describes the average hourly earnings or wage rates (the dependent variable used in this application) of 753 married white women between the ages of 30 and 60, with 428 working at some time during the year 1975. For those who did not work in 1975, the wage rate is zero,

so the variable can be classified as censored-uncensored, i.e., it follows Equation (2.1.12) with $\tau_i = 0$ for $i = 1, \ldots, 753$. This dataset presents left censored observations, since we can only observe its real value if a woman worked for pay during 1975. Thus, our purpose is to model the wage rate as a function of a set of control variables such as the wife's age (x_{1_i}), her years of schooling (x_{2_i}), husband's hours worked (x_{3_i}), husband's wage in dollars (x_{4_i}), tax rate faced by the wife (x_{5_i}), number of children younger than six years old in the household (x_{6_i}) and number of children between the ages of six and nineteen (x_{7_i}). We also consider a nonlinear relation between the wage rates and the number of years the wife worked since age 18 (see Figure 3.7). The dataset was analyzed previously by Arellano-Valle et al. (2012), Castro et al. (2014) and Massuia et al. (2015).

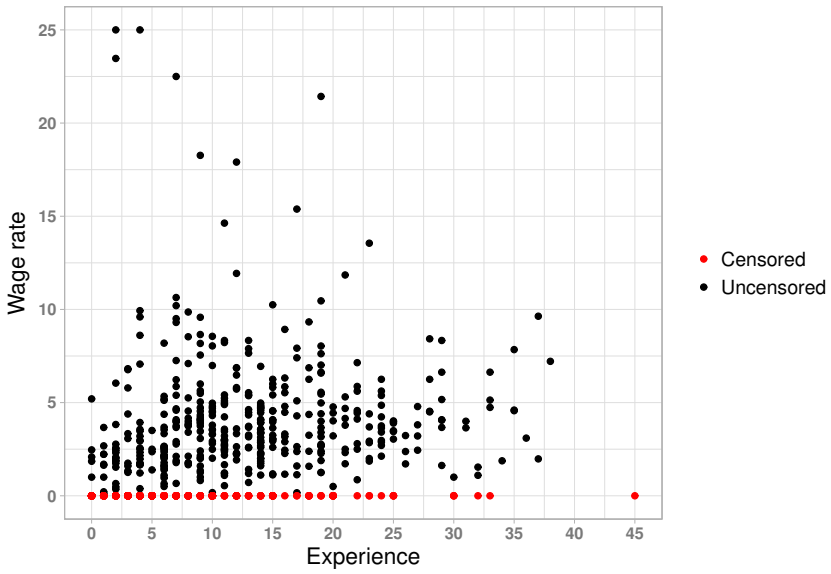

Figure 3.7: PSID-1975 dataset. Wage rates vs. number of years the wife worked (Experience).

3.2.1 Analyses of the fitted models

The data were analyzed using the SMN-PCR models considering the Student-t (T-PCR), slash (SL-PCR), contaminated normal (CN-PCR), and of course the normal distribution (N-PCR) for comparative purposes. Table 3.4 contains the MPL estimates of the parameters for the four fitted models, together with their corresponding standard errors calculated via the observed information matrix as presented in Subsection 2.1.4. Note from this table that the estimated values of ν are small, indicating a heavy-tailed behavior and consequently the lack of adequacy of the N-PCR model for this dataset. Further, we compare the results among the SMN-PCR models using the AIC value defined in Subsection 2.1.3 and the log-likelihood values $(\ell(\widehat{\boldsymbol{\theta}}))$. As expected, we can see that the SMN distributions with heavy tails have a better performance compared with the normal one, evidencing once again a clear departure from the normality assumption, with the SL-PCR model significantly better. Regarding the smoothing parameter α, there is no difference between the models with heavy tails, but in the N-PCR model the value of α is higher than the other ones. Finally, the values of the respective standard errors (SE) of the heavy tails models are smaller than those under the normal assumption,

indicating that the SMN-PCR models produce more precise estimates. Table A.1 of Appendix A presents the estimates for fitting to the simple censored regression model considering SMN distributions for the error, the SMN-CR models (see Garay et al. (2017) for more details). According to the AIC and $\ell(\widehat{\boldsymbol{\theta}})$ values, the SMN-PCR models perform better than the SMN-CR models, which indicates that including the nonparametric component improves the settings for this dataset.

Table 3.4: PSID-1975 dataset. Parameter estimates and standard errors (SE) for the SMN-PCR models.

Parameter	**Model**							
	N-PCR		**T-PCR**		**SL-PCR**		**CN-PCR**	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	0.7688	(0.0881)	0.6672	(0.0783)	0.6620	(0.0757)	0.6682	(0.0730)
β_2	-0.0634	(0.0269)	-0.0743	(0.0204)	-0.0745	(0.0201)	-0.0746	(0.0207)
β_3	-0.0008	(0.0004)	-0.0004	(0.0003)	-0.0004	(0.0003)	-0.0005	(0.0003)
β_4	-0.1801	(0.0742)	-0.1321	(0.0642)	-0.1403	(0.0622)	-0.1527	(0.0605)
β_5	-8.6192	(3.8726)	-5.7045	(3.3936)	-6.0435	(3.2619)	-6.3471	(3.1808)
β_6	-1.8132	(0.4056)	-1.7666	(0.3136)	-1.7829	(0.3103)	-1.7874	(0.3254)
β_7	0.3257	(0.1456)	0.1986	(0.1136)	0.2038	(0.1098)	0.2145	(0.1126)
σ^2	15.8854	(1.2320)	5.4806	(0.6375)	3.4223	(0.3788)	7.2270	(0.7089)
ν	-	-	2.8655	-	1.1248	-	-	-
φ	-	-	-	-	-	-	0.1	-
γ	-	-	-	-	-	-	0.1	-
$\ell(\widehat{\boldsymbol{\theta}})$	-1378.2	-	-1305.8	-	-1303.3	-	-1303.8	-
AIC	2852.2	-	2709.6	-	2704.5	-	2707.6	-

3.2.2 Diagnostics analysis

For the purpose of identifying possible observations that can affect the MPL estimates, we use the diagnostic measures presented in Section 2.2 for the PSID-1975 dataset. From the results of the ECME algorithm, Figure 3.8 shows the estimated weights $u_i = \xi_{0_i}(\widehat{\boldsymbol{\theta}})$, $i = 1, \ldots, 753$, versus the Mahalanobis distance, which is defined by $d_i^2 = (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} - \mathbf{n}_i^\top \widehat{\mathbf{f}})^2/\widehat{\sigma^2}$. For the normal case, we have that $u_i = 1$, $\forall i$ (segmented red lines). We can observe from this figure that u_i is inversely proportional to d_i^2, i.e., large d_i^2 values imply smaller u_i weights. Hence, using distributions with heavier tails than the normal leads to smaller weights being attributed to possible influential observations (see also Figure A.6, Appendix A).

To identify influential observations in a global context and following the approach described in Subsection 2.2.1, the index plots for the approximate generalized Cook's distance GD_i^1 are shown in Figure 3.9. High values of GD_i^1 suggest that the i-th observation has an impact on the MPL estimates. We can note that, women #185, #210, #349, #357, #366, #369, #394, #408 and #692 are potentially influential in the MPL estimates under the N-PCR (panel a), but for distributions with heavy tails these women are no longer influential (panels b to d). Comparing Figures 3.8 and 3.9, we see that women

who were considered influential for the normal case obtained small weights in the Student-t, slash and contaminated normal cases, as expected.

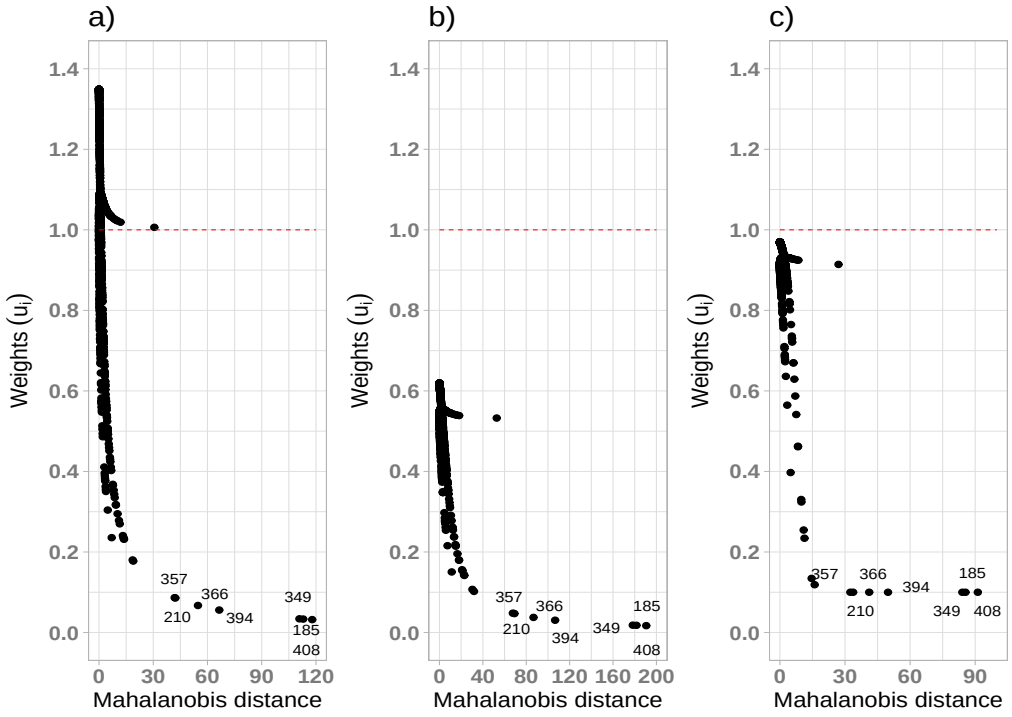

Figure 3.8: PSID-1975 dataset. Estimated weights u_i vs Mahalanobis distance d_i^2 for: a) T-PCR, b) SL-PCR and c) CN-PCR models, respectively.

Next, we study the local influence based on $M(0)$ from the conformal normal curvature $B_{f_Q,\boldsymbol{h}_l}(\boldsymbol{\theta})$ defined in Equation (2.2.5), considering the four perturbation schemes: case-weight perturbation, scale perturbation, explanatory variable perturbation and response variable perturbation. We compute the matrix $\Delta_{\boldsymbol{\theta},\boldsymbol{\omega}_0}$ for each perturbation scheme to analyze the respective local influence measures for $\widehat{\boldsymbol{\theta}}$ obtained from the best fitted model, that is the SL-PCR model. As benchmark, we use the criterion $M(0)_l > \overline{M(0)} + c^* SM(0)$, with $c^* = 4$, to classify i-th observation as potentially influential.

Examining Figure 3.10, we have that for the three first perturbation schemes (see panels a to f) women #185, #210, #349, #357, #366, #369, #394, #408 and #692 in the N-PCR model are considered influential. In addition, it is noteworthy that observations that were considered influential in the case-weight perturbation were also found influential with scale perturbation. No observation has a significant influence on the MPL estimates under the SL-PCR model, indicating the robustness of the MPL estimates against potentially influential observations. However, for the response variable perturbation (panels g to h), it was found that women #27, #55, #57, #87, #271, #298, #397 and #598 have a moderate influence in both models (normal and slash) (see Figure A.7, Appendix A). We infer that this is because the "*experience*" values in those cases are high, which naturally leads to a moderate effect on the estimates of $M(0)_l$. For comparison, we use normal benchmarks for all local influence graphs. The influence analyses for the remaining fitted models are shown in Figure A.8 given

in Appendix A.

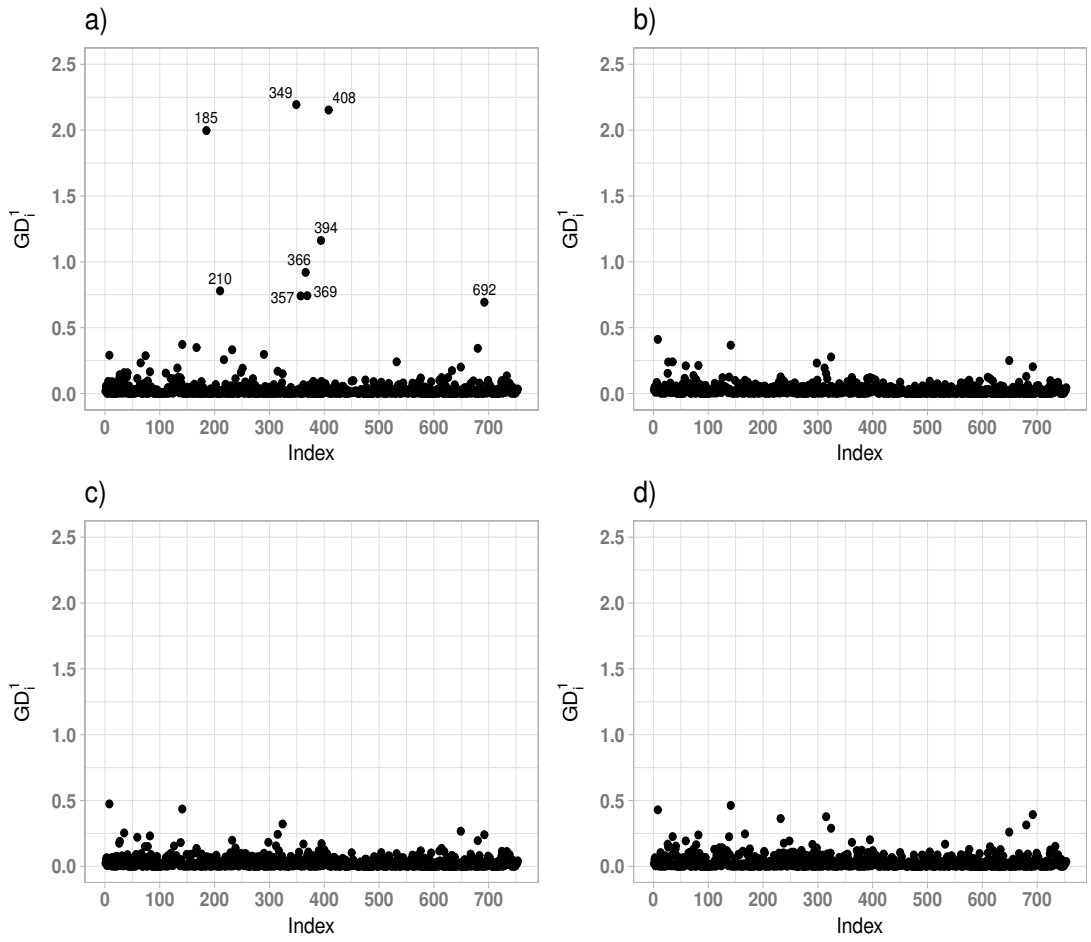

Figure 3.9: PSID-1975 dataset. Approximate generalized Cook's distance GD_i^1. a) N-PCR, b) T-PCR, c) SL-PCR and d) CN-PCR models, respectively.

3.2.3 Relative change in the MPL estimates

To detect the impact of the observations identified as potentially influential on the MPL estimates, we define the relative change (RC) as

$$RC_j(\widehat{\boldsymbol{\theta}}) = \left| \frac{\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{[-\mathcal{I}_j]}}{\widehat{\boldsymbol{\theta}}} \right| \times 100\%,$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\sigma^2})$, and $\widehat{\boldsymbol{\theta}}_{[-\mathcal{I}_j]}$ is the estimate of the parameters after the observations indexed by the set $\mathcal{I}_j$ have been removed. From all possible combinations, we drop influential cases, one by one and from all observations at the same time. Then, the sets of interest are $\mathcal{I}_j = \{e_j\}$, for $j = 1, \ldots, J$ and $\mathcal{I}_{J+1} = E$, where E is the set of all J influential case indexes, i.e. $E = \{185, 210, 349, 357, 366, 369, 394, 408, 692\}$.

For our example, Table 3.5 presents the relative change of the MPL estimates after removing the observations indexed by $\mathcal{I}_j$ and refitting the N-PCR and SL-PCR models, respectively. Note that the biggest changes occur in the N-PCR model, particularly for the parameter σ^2. As expected, the results indicate smaller changes in the MPL estimates under SL-PCR model, confirming the robust aspects when distributions with heavier tails than the normal one are used. Although some relative changes are significant, particularly with the normal distribution, most of times the SL-PCR model presented much smaller RCs. In addition, the RC for the set of observations identified as influential under the response variable perturbation scheme $E^* = \{27, 55, 57, 87, 271, 298, 397, 598\}$ were small and quite similar in both models (see also Table A.2 of Appendix A).

Table 3.5: PSID-1975 dataset. Relative change (%) of maximum penalized likelihood estimates of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ in N-PCR and SL-PCR models.

		Parameter							
Model	Dropped	$RC_{\widehat{\beta}_1}$	$RC_{\widehat{\beta}_2}$	$RC_{\widehat{\beta}_3}$	$RC_{\widehat{\beta}_4}$	$RC_{\widehat{\beta}_5}$	$RC_{\widehat{\beta}_6}$	$RC_{\widehat{\beta}_7}$	$RC_{\widehat{\sigma^2}}$
	185	1.908	7.551	11.70	7.446	2.109	6.693	4.575	8.718
	210	0.855	2.244	3.378	0.163	0.216	3.675	6.568	3.235
	349	1.218	20.28	6.792	2.789	7.210	0.485	12.48	8.291
	357	2.260	9.032	4.944	4.104	0.002	3.061	6.400	3.677
N-PCR	366	0.030	14.11	5.085	11.98	4.974	0.348	4.735	4.481
	369	0.425	8.584	5.543	13.18	2.333	3.420	3.465	1.614
	394	3.525	8.433	3.675	2.474	0.257	2.077	2.860	5.467
	408	0.335	3.442	2.979	1.401	5.536	7.793	13.77	9.524
	692	0.709	0.713	1.587	1.687	0.390	0.231	0.921	0.576
	E	11.24	1.558	19.32	6.239	14.59	10.10	10.33	46.95
	185	0.015	0.182	7.375	1.812	0.923	0.280	0.537	4.376
	210	0.138	0.179	6.497	0.733	0.421	0.564	1.806	2.526
	349	0.047	1.004	7.762	1.114	0.277	0.026	0.163	4.353
	357	0.396	1.195	8.373	1.625	0.503	0.458	2.160	2.235
SL-PCR	366	0.137	1.293	8.811	0.662	0.130	0.062	1.466	2.854
	369	0.262	2.375	9.181	5.238	1.036	1.063	2.470	0.883
	394	0.355	0.678	8.331	0.735	0.566	0.309	1.166	3.046
	408	0.073	0.257	8.894	1.216	0.412	0.468	0.205	4.542
	692	0.358	0.213	5.980	1.120	0.103	0.044	0.534	0.355
	E	1.445	8.358	42.01	25.55	12.68	5.076	24.97	10.29

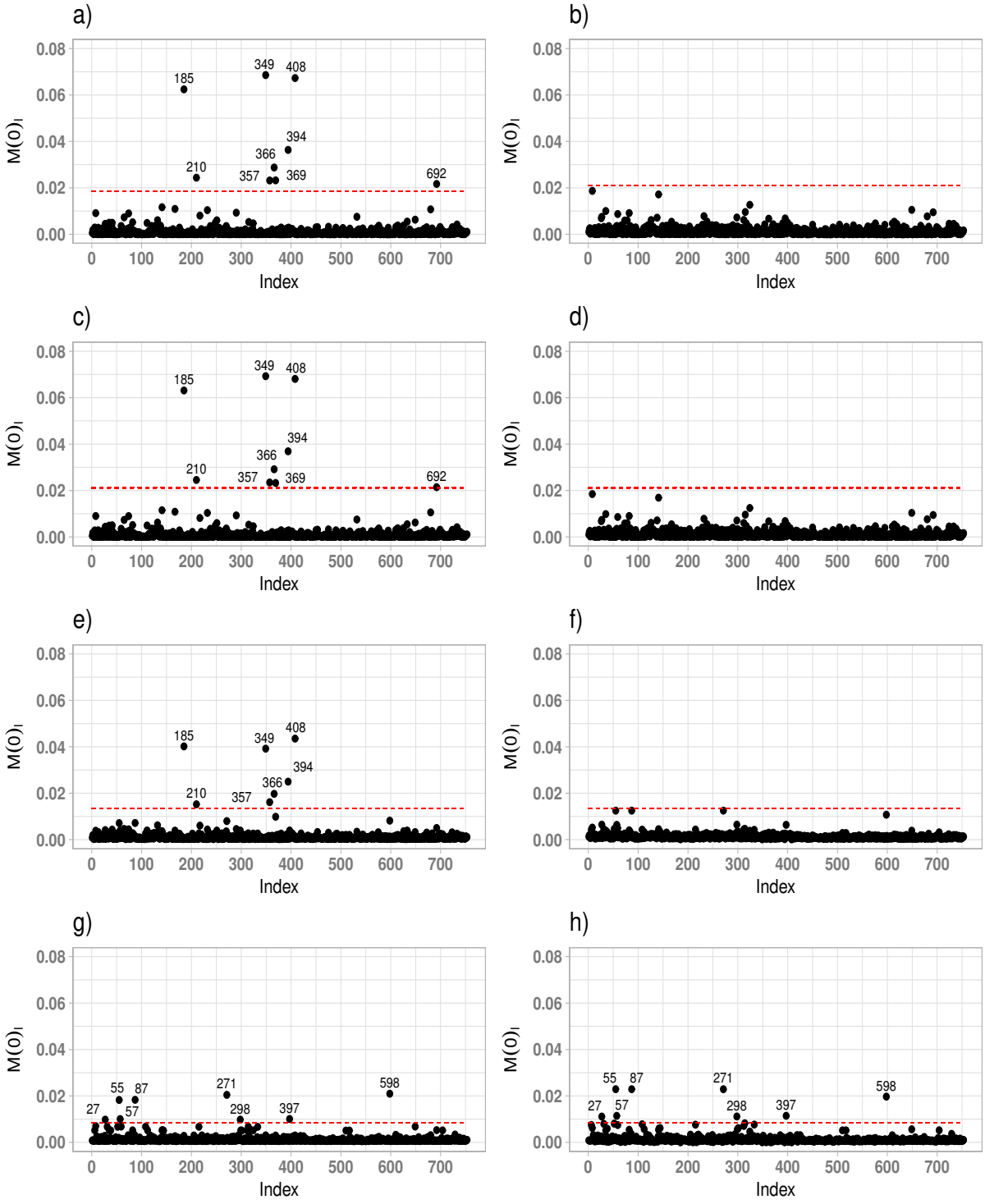

Figure 3.10: PSID-1975 dataset. Index plots of $M(0)_l$ for assessing local influence. Different perturbations schemes (case-weight, scale, explanatory variable and response variable perturbation) are shown in the rows from top to bottom. The N-PCR and SL-PCR models correspond to the columns from left to right.

Chapter 4

Concluding remarks

4.1 Technical production

In this section we will describe the technical production developed in this dissertation.

4.1.1 Submitted paper

- **Title**: "Estimation and diagnostics for partially linear censored regression models based on heavy-tailed distributions".
  Authors: Marcela N. Lemus, Victor H. Lachos, Larissa A. Matos and Christian E. Galarza.

4.1.2 R package

`PartCensReg` : Partially linear censored regression models Based on Heavy-Tailed Distributions.

It estimates the parameters of a partially linear regression censored model via maximum penalized likelihood through of ECME algorithm. The model belong to the semiparametric class, that including a parametric and nonparametric component. The error term considered belongs to the scale-mixture of normal (SMN) distribution, that includes well-known heavy tails distributions as the Student-t distribution, among others. To examine the performance of the fitted model, case-deletion and local influence techniques are provided to show its robust aspect against outlying and influential observations. This work is based in Ferreira and Paula (2017) but considering the SMN family.

Description

The principal function of this package is the `Cens.SMN.PCR`, which return the MPL estimates obtained through of ECME algorithm for partially linear regression models with censored data under scale-mixture of normal (SMN) distributions (some members are the normal, Student-t, slash and contaminated normal distribution). The types of censoring considered are left and right. Graphics for diagnostic analysis such as case-deletion and local influence techniques are provided to show its robust aspect against outlying and influential observations:

R Code

```
Cens.SMN.PCR(x,y,c,cens = "left",tt,nu = NULL,error = 10^-6,
iter.max = 200,type = "Normal",alpha.FIX = TRUE,nu.FIX = TRUE,
alpha.in = 10^-3,k = 1,Diagnostic = TRUE,a = 2)
```

Arguments

`x`	Matrix or vector of covariates.
`c`	Vector of censoring indicators. For each observation: 1 if censored and 0 if non-censored.
`y`	Vector of responses.
`cens`	`'left'` for left censoring and `'right'` for rigth censoring.
`tt`	Vector of values of a continuous covariate for the nonparametric component of the model.
`nu`	Initial value of the parameter of the SMN family. In the case of the Student-t and slash is a scalar, in the contaminated normal is a vector bidimensional.
`error`	The convergence maximum error. By default =`10^-6`.
`iter.max`	The maximum number of iterations of the ECME algorithm. By default =`200`.
`type`	Represents the type of distribution to be used in fitting: `'Normal'` for normal, `'T'` for Student-t, `'Slash'` for slash and `'NormalC'` for contaminated normal distribution respectively. By default =`'Normal'`.
`alpha.FIX`	`TRUE` or `FALSE`. Indicate if smoothing parameter will be estimated. By default =`TRUE`.
`nu.FIX`	`TRUE` or `FALSE`. Indicate if ν will be estimated. By default =`TRUE`.
`alpha.in`	Initial value of smoothing parameter.
`k`	For the local influence in explanatory variable perturbation, indicates the k-th explanatory variable (assumed continuous) of the design matrix X to be perturbed.
`Diagnostic`	`TRUE` or `FALSE`. Indicates if diagnostic graph should be built for the fitted model (index plot in local influence). By default =`TRUE`.
`a`	The value for a considered in the benchmark value for the index plot in local influence: $M(0)_l > \bar{M(0)} + a * SM(0)$.

Details

We consider a partial linear model which belongs to the class of semiparametric regression models with vector of response $Y = (Y_1, ..., Y_n)$ and with errors ϵ_i which are independent and identically distributed according to a SMN distribution. To be more precise,

$$Y_i = x_i^T \beta + n_i^T f + \epsilon_i,$$

for $i = 1, ..., n$, where $f = (f(t_1^0), ..., f(t_r^n)^T$ is an $r \times 1$ vector with $t_1^0, ..., t_r^n$ being the distinct and ordered values of t_i; n_i is a $r \times 1$ vector of incidence whose s-th element equals the indicator function $I(t_i = t_s^0)$ for $s = 1, ..., r$.

Value

`beta` ECME estimates for the parametric component.

`sigma2` ECME estimates for the scale parameter.

`Alpha` If `alpha.FIX=FALSE`, it returns the estimated value of the smoothing parameter, else returns the initial value assigned in `alpha.in`.

`AIC` AIC criteria for model selection.

`ff` ECME estimates for the nonparametric component.

`yest` Predicted values of the model.

`loglik` Value of the log-likelihood under the fitted model.

`iter` Number of iterations of the ECME algorithm.

`nu` If `nu.FIX=FALSE`, it returns the estimated value of ν parameter, else returns the initial value assigned in `nu`.

`MI` Observed information matrix.

`D` A list of objects for diagnostic analysis that contains: the Hessian matrix (`Hessian`), values for generalized Cook's distance (`GD1`) and the values of the conformal normal curvature for the following perturbation schemes: Case-weight (`Curvature_W`), scale (`Curvature_S`), explanatory variable (`Curvature_E`) and response variable (`Curvature_R`).

Observation 1. The package estimates the value $\boldsymbol{\nu}$ in each iteration taking as an estimate the argument that maximizes the actual marginal log-likelihood function, already evaluated in the estimates of $\boldsymbol{\beta}$ and σ^2. However, the diagnostic analysis is performed considering the estimated final value of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\nu}$ obtained in the last iteration of the ECME algorithm.

Example

R Code

```
dtawage = get(data(PSID1976,package = "AER"))
y  = dtawage$wage
cc = c(rep(0,428),rep(1,325))
tt = dtawage$exper
x  = cbind(dtawage$education, dtawage$age, dtawage$hhours,
dtawage$hwage, dtawage$tax, dtawage$youngkids, dtawage$oldkids)

#Normal case by default with only 10 iterations
PCR.default1 = Cens.SMN.PCR(x=x, y=y, c=cc, cens="left",tt =tt,
iter.max = 10,Diagnostic = FALSE)
```

4.2 Conclusion

In this work, we propose a maximum penalized likelihood implementation of a robust alternative to the partially linear regression (PCR) model with censored response, where the normal distribution of the random terms are replaced by a scale mixture of normal (SMN) distribution, which allow to work with censored data and outliers. Special cases of the SMN distributions are Student-t, slash, contaminated normal, normal, among others. Thus, this work generalizes the papers of Garay et al. (2017) and Massuia et al. (2015) by incorporating a nonparametric component into the model that permits an easy and flexible modeling of possible nonlinear pattern introduced by some covariate.

In Chapter 2, from a frequentist perspective, we develop SMN-PCR model, where the stochastic representation of the model allows a simple implementation of an EM-type algorithm (ECME algorithm) to get the MPL estimates. Expressions for the standard errors approximation were derived from the inverse of the observed information matrix. We also propose influence diagnostic tools for detecting influential observations in the context of PCR models with heavy-tailed distribution errors. The diagnostic analysis is based on local influence techniques presented in Zhu and Lee (2001) and Zhu et al. (2001).

Moreover, in Chapter 3, we carried out extensive simulation studies and an application to real dataset to evaluate the performance of the proposed methodology. The results showed that the partially linear regression model with censored response under scale mixtures of normal distribution errors is very robust against outlying observations, outperforming traditional normal errors model. We also use simulation to investigate asymptotic properties of the parameter estimates, where we could observe that for large sample the MPL estimates has good asymptotic properties, i.e. the bias and MSE tends to zero. Besides,

in the estimation of the nonparametric component the variability among estimates decreases when the sample size increases. Further, we applied our method to a real dataset to illustrate how the procedure developed can be used to evaluate model assumptions, identify outliers and obtain robust parameter estimates. To the our knowledge, this work provides a first attempt to incorporate censoring in the context of partially linear models with heavy-tailed distributions from a likelihood-based perspective. The package `PartCensReg` (Lemus et al., 2018) give computational support for estimation procedure and diagnostic analysis. The package is available in the CRAN repository.

4.3 Future research

A natural extension would be to incorporate skewness and heavy tailedness simultaneously using scale mixtures of skew-normal (SMSN) distributions, as proposed in Lachos et al. (2010). Other extensions of the current work include considering semiparametric mixed effects models with censored data, following the same lines of ideas proposed by Matos et al. (2013) and Matos et al. (2015). Finally, extend the analysis of local influence to a subset of the parameters of interest (local influence analysis on sub-vectors), following the work of Zhu et al. (2003), Ibacache-Pulgar and Paula (2011), Chen et al. (2012) and Relvas and Paula (2016).

Bibliography

Andrews, D. F. and C. L. Mallows (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B 36*, 99–102.

Arellano-Valle, R. B., L. M. Castro, G. González-Farías, and K. A. Muñoz-Gajardo (2012). Student-t censored regression model: properties and inference. *Statistical Methods & Applications 21*, 453–473.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*, 171–178.

Castro, L. M., V. H. Lachos, G. P. Ferreira, and R. B. Arellano-Valle (2014). Partially linear censored regression models using heavy-tailed distributions: A Bayesian approach. *Statistical Methodology 18*, 345–352.

Chen, X.-d., N.-s. Tang, and X.-r. Wang (2012). Local influence analysis for semiparametric reproductive dispersion nonlinear models. *Acta Mathematicae Applicatae Sinica, English Series 28*(1), 75–90.

Choongrak, K., U. P. Byeong, and K. Woochul (2002). Influence diagnostics in semiparametric regression models. *Statistics & Probability Letters 60*(1), 49–58.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics 19*(1), 15–18.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B 48*, 133–169.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression*. Boca Raton, FL: Chapman & Hall/CRC.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1–38.

Ferreira, C. S. and G. A. Paula (2017). Estimation and diagnostic for skew-normal partially linear models. *Journal of Applied Statistics 44*(16), 3033–3053.

Garay, A. M., H. Bolfarine, V. H. Lachos, and C. R. B. Cabral (2015). Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. *Journal of Applied Statistics 42*, 2694–2714.

Garay, A. M., V. H. Lachos, H. Bolfarine, and C. R. B. Cabral (2017). Linear censored regression models with scale mixtures of normal distributions. *Statistical Papers 58*, 247–278.

Genç, A. İ. (2013). Moments of truncated normal/independent distributions. *Statistical Papers 54*, 741–764.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259.

Green, P. J. and B. W. Silverman (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. CRC Press.

Hadi, E. (2016). Local influence in ridge semiparametric models. *Journal of Statistical Computation and Simulation 86*(17), 3357–3370.

Härdle, W., M. Müller, S. Sperlich, and A. W. Werwatz (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag Berlin Heidelberg.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Wiley Online Library.

Heckman, E. N. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(2), 413–436.

Holland, D. A. (2017). Penalized spline estimation in the partially linear model. *Journal of Multivariate Analysis 153*, 211–235.

Ibacache-Pulgar, G. and G. A. Paula (2011). Local influence for student-t partially linear models. *Computational Statistics & Data Analysis 55*(3), 1462–1478.

Ibacache-Pulgar, G., G. A. Paula, and F. Cysneiros (2013). Semiparametric additive models under symmetric distributions. *Test 22*, 103–121.

Lachos, V. H., T. Angolini, and C. A. Abanto-Valle (2011). On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. *Statistical Papers 52*(3), 567–590.

Lachos, V. H., P. Ghosh, and R. B. Arellano-Valle (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica 20*. SS-08-045.

Lee, S. Y. and L. Xu (2004). Influence analysis of nonlinear mixed-effects models. *Computational Statistics and Data Analysis 45*, 321–341.

Lemus, M. N., C. E. Galarza, L. A. Matos, and V. H. Lachos (2018). *PartCensReg: Partially Censored Regression Models Based on Heavy-Tailed Distributions*. R package version 1.38.

Liang, H. (2006). Estimation in partially linear models and numerical comparisons. *Computational Statistics & Data Analysis 50*(3), 675–687.

Liu, C. and D. B. Rubin (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika 80*, 267–278.

Mark, R. S., P. Bacchetti, and N. P. Jewell (1994). Variances for maximum penalized likelihood estimates obtained via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(2), 345–352.

Massuia, B. M., C. R. B. Cabral, L. A. Matos, and V. H. Lachos (2015). Influence diagnostics for student-t censored linear regression models. *Statistics 49*(5), 1074–1094.

Matos, L. A., D. Bandyopadhyay, L. M. Castro, and V. H. Lachos (2015). Influence assessment in censored mixed-effects models using the multivariate student's-t distribution. *Journal of multivariate analysis 141*, 104–117.

Matos, L. A., V. H. Lachos, N. Balakrishnan, and F. V. Labra (2013). Influence diagnostics in linear and nonlinear mixed-effects models with censored data. *Computational Statistics & Data Analysis 57*(1), 450–464.

McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions*. New Jersey: John Wiley & Sons.

Meng, X. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika 81*, 633–648.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, 765–799.

Osorio, F., G. A. Paula, and M. Galea (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics and Data Analysis 51*, 4354–4368.

Poon, W.-Y. and Y. S. Poon (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 51–61.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Relvas, C. E. M. and G. A. Paula (2016). Partially linear models with first-order autoregressive symmetric errors. *Statistical Papers 57*(3), 795–825.

Ruppert, M. D., P. Wand, and J. R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological) 50*(3), 413–436.

Vanegas, L. H. and G. A. Paula (2017). Log-symmetric regression models under the presence of non-informative left-or right-censored observations. *TEST 26*(2), 405–428.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics 11*(1), 95–103.

Wu, L. (2010). *Mixed Effects Models for Complex Data*. Boca Raton, FL: Chapman & Hall/CRC.

Zhu, H. and S. Lee (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B 63*, 111–126.

Zhu, H., S.-Y. Lee, B.-C. Wei, and J. Zhou (2001). Case-deletion measures for models with incomplete data. *Biometrika 88*(3), 727–737.

Zhu, Z.-Y., X. He, and W.-K. Fung (2003). Local influence analysis for penalized gaussian likelihood estimators in partially linear models. *Scandinavian Journal of Statistics 30*(4), 767–780.

Appendix A

Supplementary material for Chapter 3

We include here supplementary material of Chapter 3. Section A.1 presents the behavior of the 500 MC samples of the nonparametric component for the N-PCR, SL-PCR and CN-PCR models respectively. Also Figures A.4 and A.5 show the behavior of the diagnostic measures through index plots for scale perturbation and response variable perturbation for the local influence approach described in the Subsection 2.2.2 of Chapter 2. On the other hand, Section A.2 reports additional results of the wage rate dataset of the Section 3.2, where some Figures in diagnostic analysis and relative change in the estimates for observations identified as potentially influential in the response variable perturbation are presented.

A.1 Simulation study

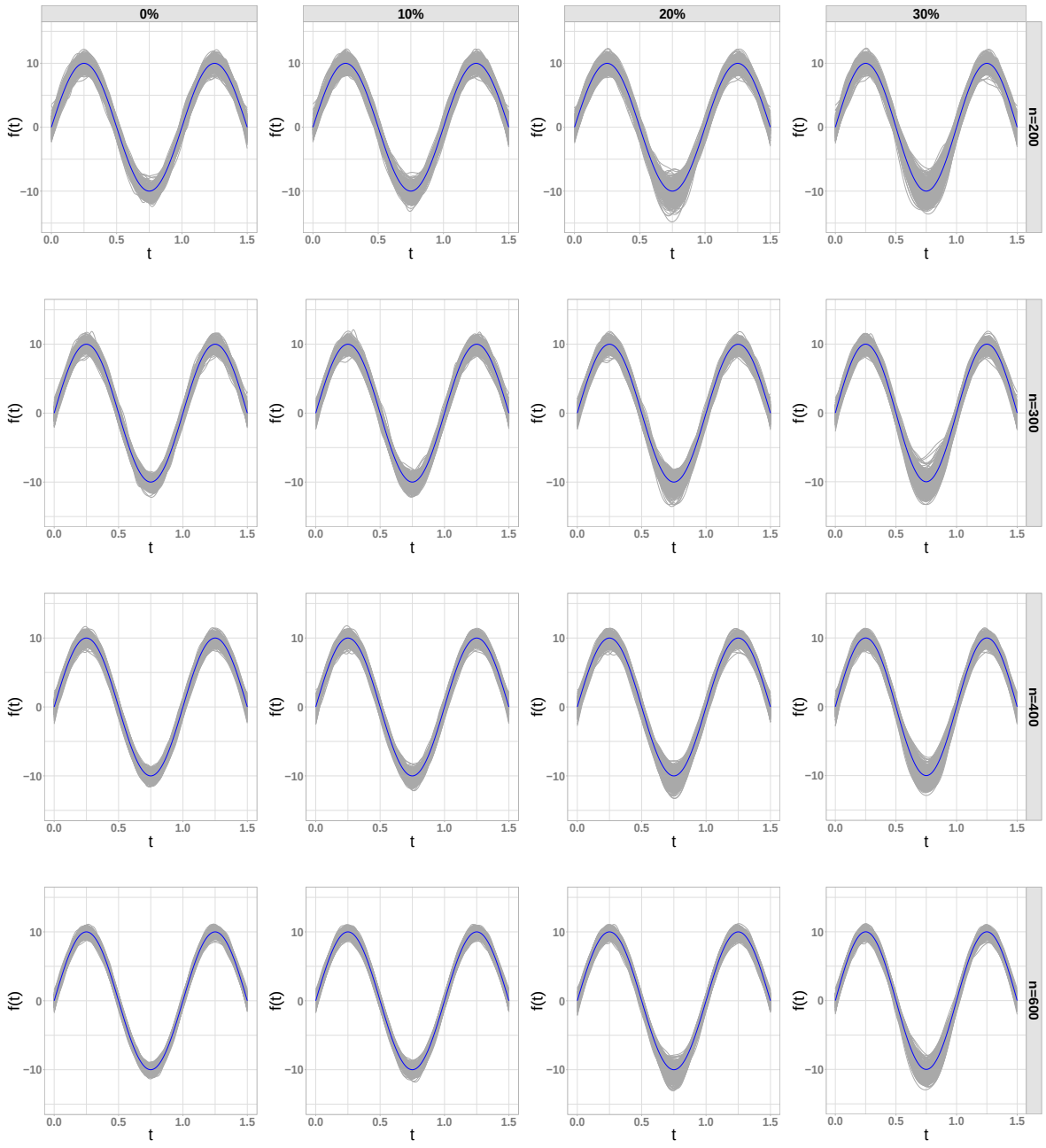

Figure A.1: Simulated data. Behavior of the nonparametric component based on 500 samples from the N-PCR model. True curve (blue line) and adjusted curves (gray lines).

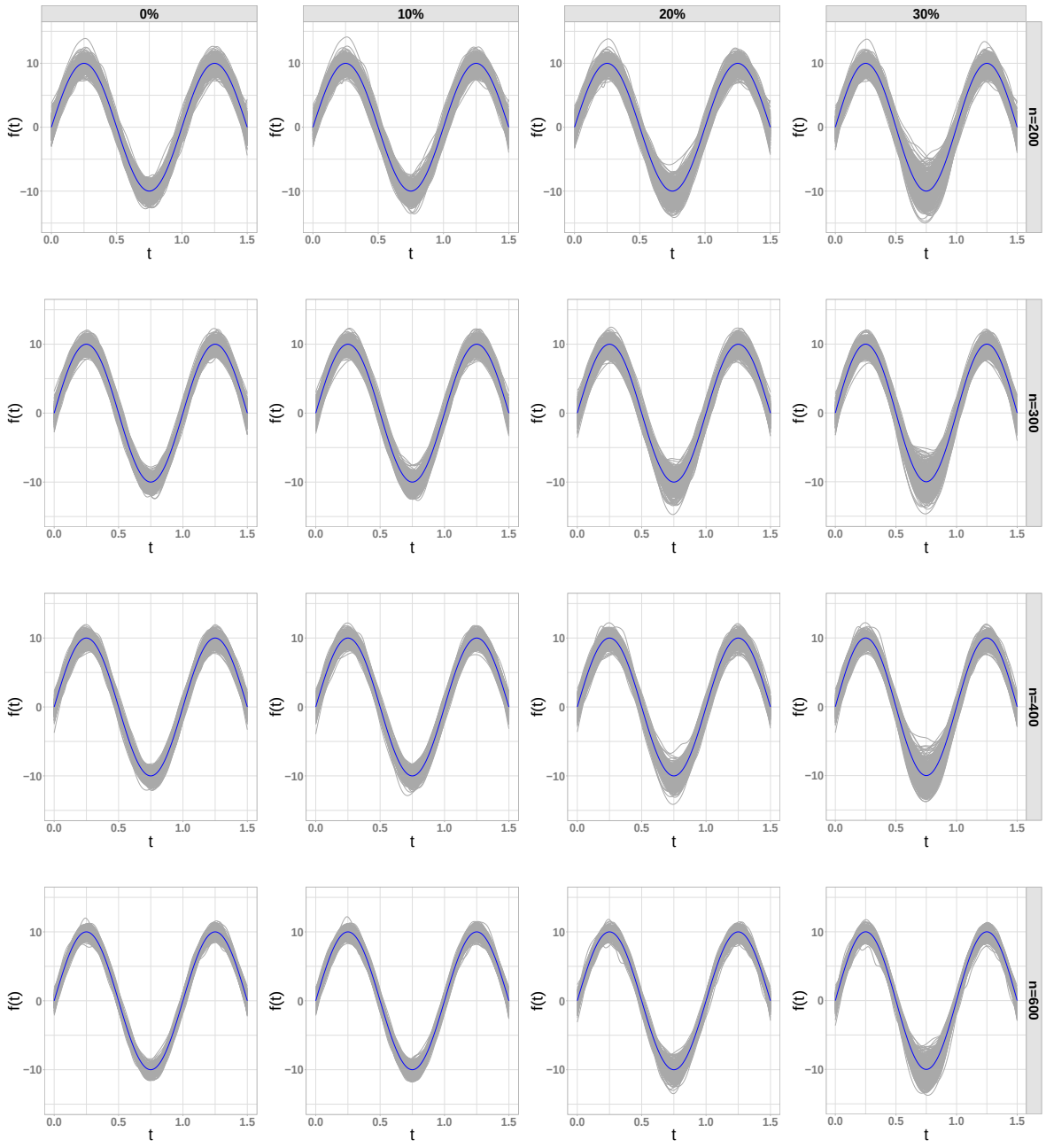

Figure A.2: Simulated data. Behavior of the nonparametric component based on 500 samples from the SL-PCR model. True curve (blue line) and adjusted curves (gray lines).

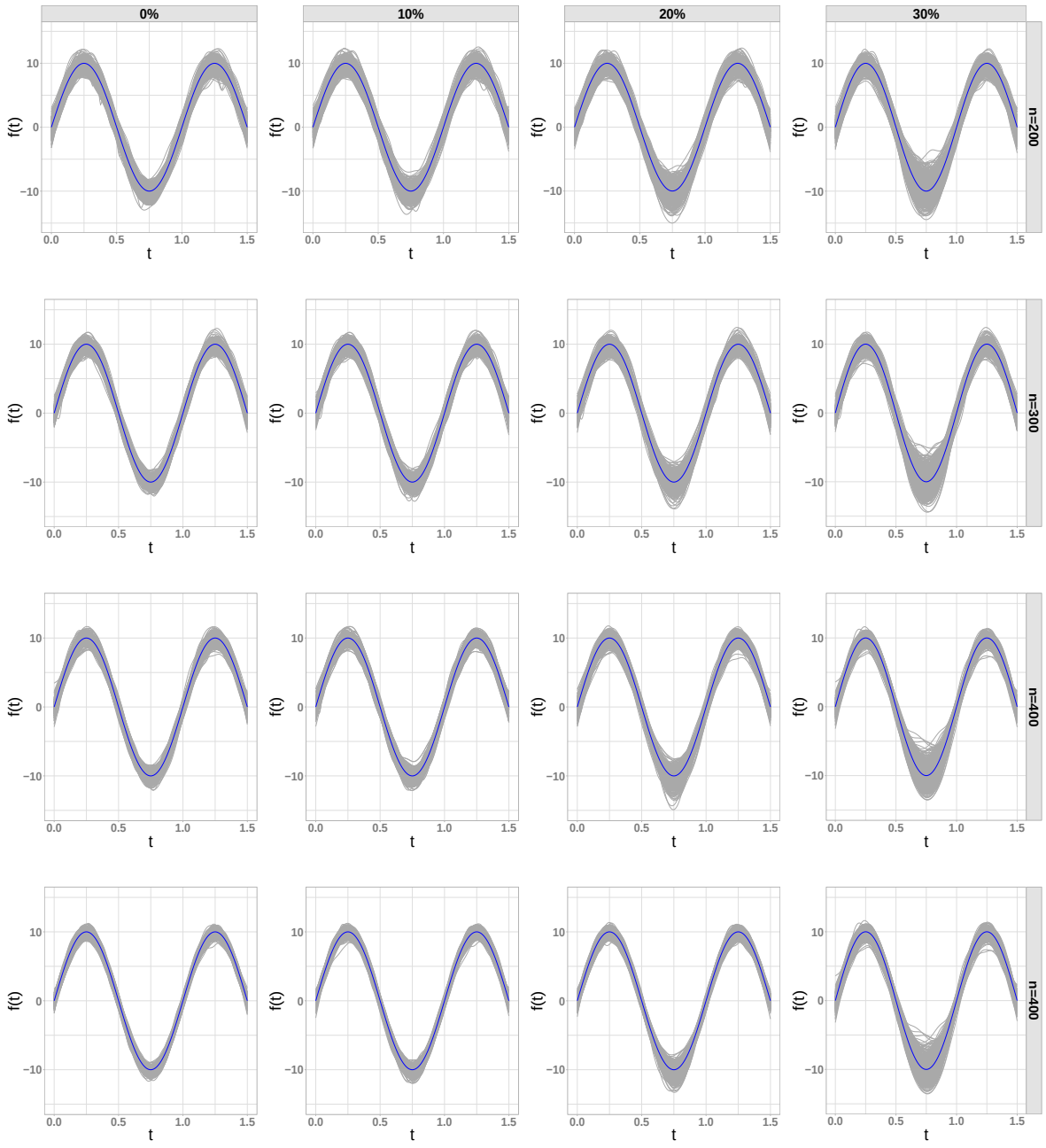

Figure A.3: Simulated data. Behavior of the nonparametric component based on 500 samples from the CN-PCR model. True curve (blue line) and adjusted curves (gray lines).

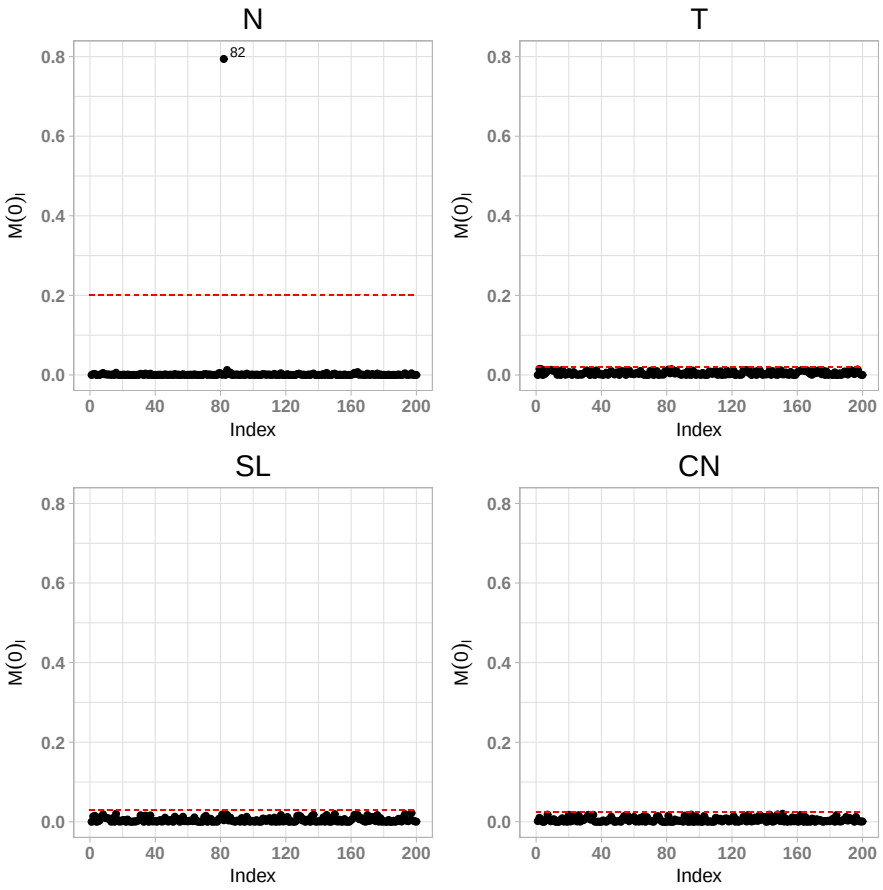

Figure A.4: Simulated data. Index plot of $M(0)_l$ for assessing local influence for contamination rate 8β. Scale perturbation (simulation No.3).

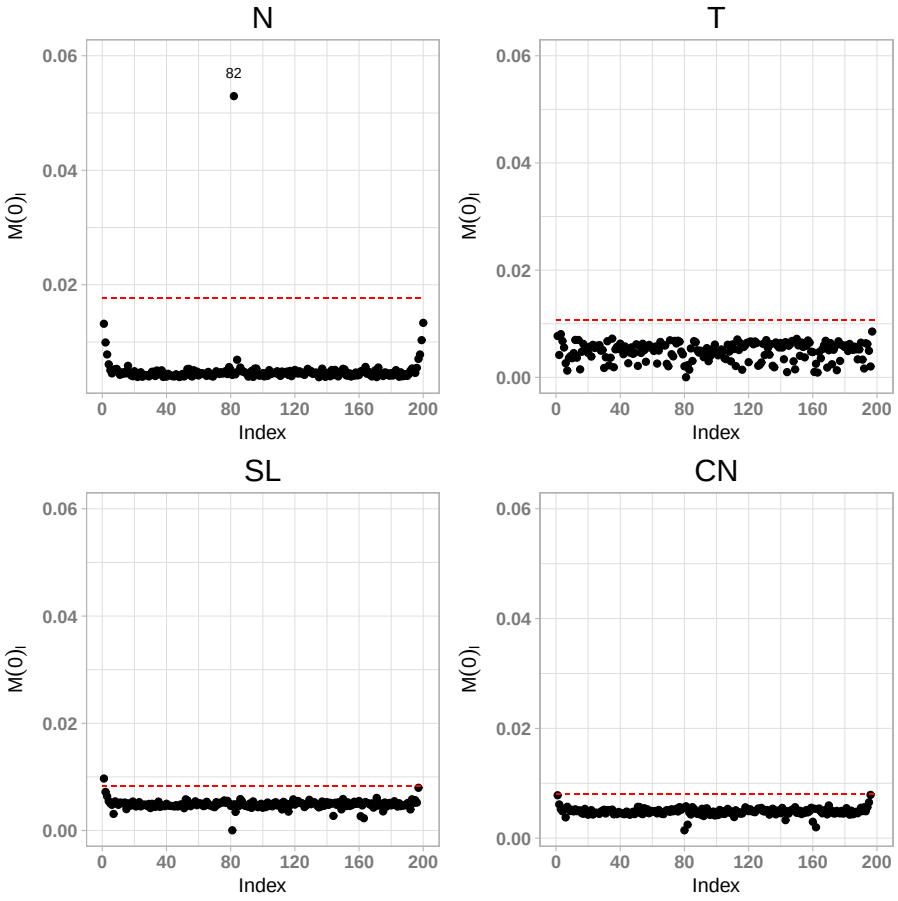

Figure A.5: Simulated data. Index plot of $M(0)_l$ for assessing local influence for contamination rate 8β. Response variable perturbation (simulation No.3).

A.2 Application: Wage rate data

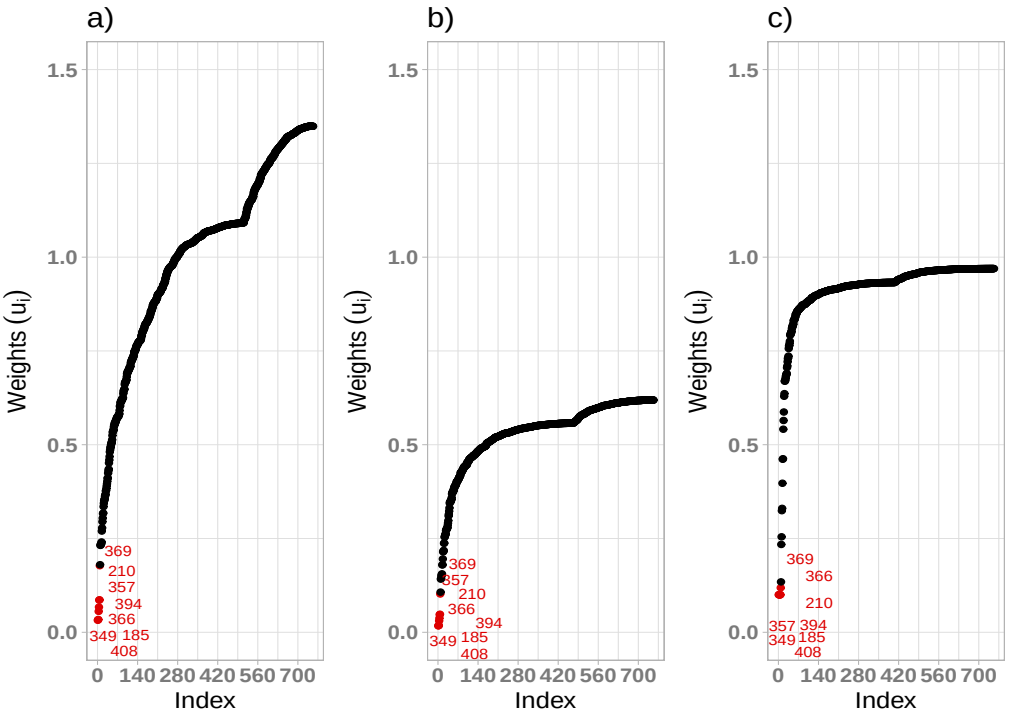

Figure A.6: PSID-1975 dataset. Estimated weights u_i for the: a) T-PCR, b) SL-PCR and c) CN-PCR models.

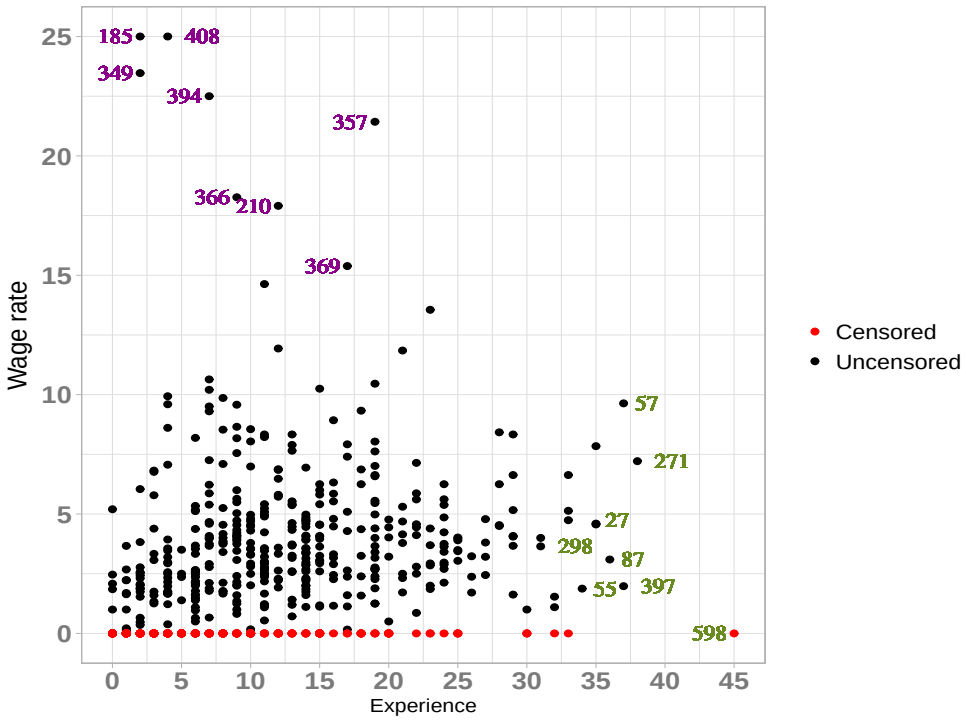

Figure A.7: PSID-1975 dataset. Potentially influential observations are numbered.

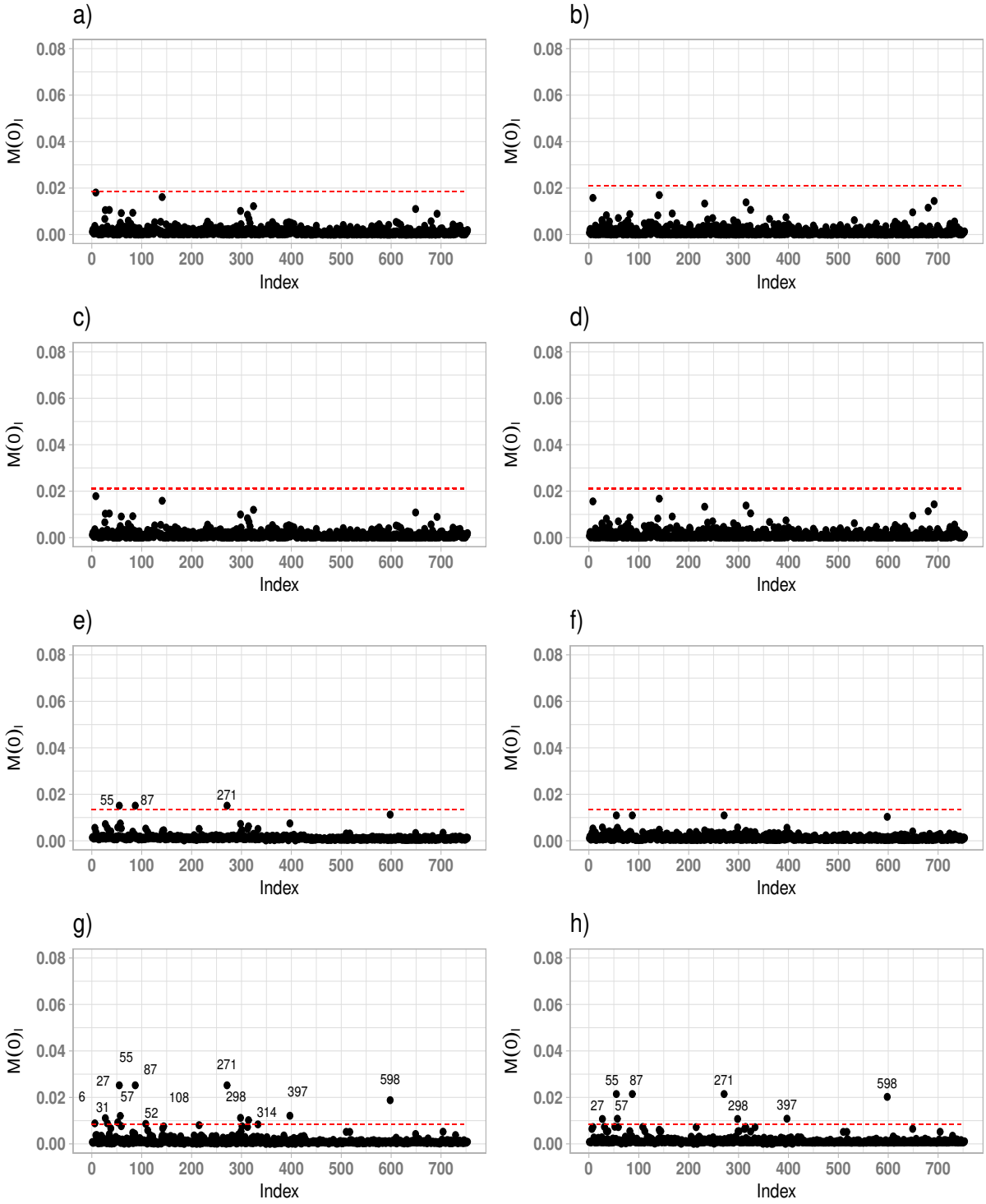

Figure A.8: PSID-1975 dataset. Index plots of $M(0)_l$ for assessing local influence. Different perturbations schemes (case-weight, scale, explanatory variable and response variable perturbation) are shown in the rows from top to bottom. The T-PCR and CN-PCR models correspond to the columns from left to right.

Table A.1: PSID-1975 dataset. Parameter estimates and standard errors (SE) for the SMN-CR models (Garay et al. (2017)).

Parameter	**Model**							
	N-CR		**T-CR**		**SL-CR**		**CN-CR**	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	0.9403	(0.0819)	0.8897	(0.0649)	0.8713	(0.0633)	0.8607	(0.0617)
β_2	-0.0271	(0.0256)	-0.0300	(0.0206)	-0.0285	(0.0206)	-0.0270	(0.0209)
β_3	-0.0010	(0.0003)	-0.0009	(0.0002)	-0.0008	(0.0002)	-0.0008	(0.0002)
β_4	-0.2330	(0.0534)	-0.2288	(0.0448)	-0.2255	(0.0449)	-0.2234	(0.0457)
β_5	-7.5911	(1.9007)	-6.2931	(1.4196)	-6.3275	(1.4247)	-6.3931	(1.4381)
β_6	-2.4520	(0.4336)	-2.5243	(0.3703)	-2.4534	(0.3662)	-2.4275	(0.3682)
β_7	0.1313	(0.1500)	0.0628	(0.1200)	0.0553	(0.1191)	0.0464	(0.1202)
σ^2	20.0093	(0.7680)	9.7277	(0.9340)	6.2229	(0.5879)	10.5138	(0.9266)
ν	-	-	4.0058	-	1.3853	-	-	-
φ	-	-	-	-	-	-	0.1	-
γ	-	-	-	-	-	-	0.1	-
$\ell(\widehat{\boldsymbol{\theta}})$	-1464.173	-	-1418.94	-	-1415.26	-	-1411.28	-
AIC	2944.345	-	2855.89	-	2848.52	-	2842.56	-

Table A.2: PSID-1975 dataset. Relative change (%) of maximum penalized likelihood estimates of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ in N-PCR and SL-PCR models, observations #27, #55, #57, #87, #271, #298, #397 and #598 and E^*.

Model	Dropped	**Parameter**							
		$RC_{\widehat{\beta}_1}$	$RC_{\widehat{\beta}_2}$	$RC_{\widehat{\beta}_3}$	$RC_{\widehat{\beta}_4}$	$RC_{\widehat{\beta}_5}$	$RC_{\widehat{\beta}_6}$	$RC_{\widehat{\beta}_7}$	$RC_{\widehat{\sigma^2}}$
N-PCR	27	0.302	0.458	3.506	0.141	0.282	0.296	0.607	0.025
	55	0.072	0.208	2.309	0.371	0.052	0.069	0.206	0.464
	57	0.276	0.226	2.331	0.316	0.093	0.037	0.071	0.133
	87	0.069	0.224	2.298	0.353	0.031	0.072	0.216	0.465
	271	0.065	0.204	2.353	0.349	0.077	0.069	0.160	0.426
	298	0.270	0.174	2.811	0.133	0.234	0.113	0.124	0.265
	397	0.064	0.197	2.518	0.001	0.082	0.076	0.131	0.193
	598	0.010	0.014	2.358	0.032	0.045	0.003	0.035	0.031
	E^*	0.354	1.318	3.218	0.757	0.769	0.611	0.371	1.960
SL-PCR	27	0.542	0.200	8.226	0.088	0.570	0.180	0.294	0.037
	55	0.044	0.025	6.968	0.228	0.216	0.059	0.206	0.643
	57	0.845	0.087	4.966	0.965	0.741	0.111	0.405	0.110
	87	0.037	0.043	6.944	0.205	0.188	0.063	0.189	0.643
	271	0.028	0.043	6.978	0.246	0.227	0.059	0.186	0.635
	298	0.520	0.190	8.151	0.119	0.511	0.179	0.289	0.044
	397	0.818	0.079	4.988	0.956	0.767	0.105	0.392	0.121
	598	0.005	0.065	6.529	0.031	0.059	0.003	0.006	0.010
	E^*	0.073	0.860	8.504	0.052	0.854	0.392	0.851	3.428

Annex I

Natural cubic splines

Smoothing splines is used in penalized least squares regression problem for find the function $g(\cdot)$ with two continuous derivatives that minimizes the penalized sum of squares

$$S(g) = \sum_{i=1}^{n}[Y_i - g(t_i)]^2 + \alpha \int_a^b [g''(x)]^2 dx, \tag{I.0.1}$$

where the first term in Equation (I.0.1) is the residual sum of squares and the second term is a roughness penalty that ensures that the curve is determined not only by its goodness-of-fit to the data, quantified by the residual sum of squares but also by its roughness $\int g''^2$. Here, $\alpha > 0$ is a smoothing parameter in which large values produce smoother curves while smaller values more wiggly curves (Hastie and Tibshirani, 1990). The aim of penalizing is to reduce the solution of the parametric space to avoid overfitting. The g function can be characterized and computed through natural cubic spline.

Definition 2. Suppose that the values of a random variable are in the interval $[a, b]$ and we have real numbers t_i satisfying $a < t_1, \ldots, t_n < b$. A function g defined on $[a, b]$ is a cubic spline if on each of the intervals $(a, t_1), (t_1, t_2), \ldots (t_n, b)$, g is a cubic polynomial and the polynomial pieces fit together at the points with and its first and second derivatives are continuous at each t_i and hence on the whole of $[a, b]$ (Green and Silverman, 1993). The points t_i are called knots and has $m = n - 2$ internal knots. Moreover, a cubic spline on an interval $[a, b]$ will be said to be a natural cubic spline (NCS) if its second and third derivatives are zero at a and b (natural boundary conditions).